Liquidity Cycles and Make/Take Fees in Electronic Markets*

Thierry Foucault
HEC School of Management, Paris
1 rue de la Liberation
78351 Jouy en Josas, France
foucault@hec.fr

Ohad Kadan
Olin Business School
Washington University in St. Louis
Campus Box 1133, 1 Brookings Dr.
St. Louis, MO 63130
kadan@wustl.edu

Eugene Kandel
School of Business Administration,
and Department of Economics,
Hebrew University,
91905, Jerusalem, Israel
mskandel@mscc.huji.ac.il

October 2009

Abstract

We develop a dynamic model of trading with two specialized sides: traders posting quotes ("market makers") and traders hitting quotes ("market takers"). Traders monitor the market to seize profit opportunities, generating high frequency make/take liquidity cycles. Monitoring decisions by market-makers and market-takers are self-reinforcing, generating multiple equilibria with differing liquidity levels and duration clustering. The trading rate is typically maximized when makers and takers are charged different fees or even paid rebates, as observed in reality. The model yields several empirical implications regarding the determinants of make/take fees, the trading rate, the bid-ask spread, and the effect of algorithmic trading on these variables. Finally, algorithmic trading can improve welfare because it increases the rate at which gains from trade are realized.

Keywords: Liquidity, Monitoring, Make/Take Fees, Duration Clustering, Algorithmic trading, Two-Sided Markets.

*We thank Torben Andersen, Hank Bessembinder, Bruno Biais, Lawrence Glosten, Jeff Harris, Lawrence Harris, Pete Kyle, Terrence Hendershott, Ernst Maug, Albert Menkveld, Marios Panayides, Sébastien Pouget, Michel Robe, Jean-Charles Rochet, Gideon Saar, Elu Von Thadden, and participants at the 2009 WFA Meeting in San Diego, the NYSE-Euronext Amsterdam-Tinbergen Institute workshop on volatility and liquidity, the FBF-IDEI conference on investment banking and financial markets in Toulouse and the Warwick Business School conference on high frequency econometrics, as well as seminar participants at the CTFC, Humboldt University, University of Mannheim, University of Toronto, and University College in Dublin for their useful comments. All errors are ours.

1 Introduction

Securities trading, especially in equities markets, increasingly takes place in electronic limit order markets. The trading process in these markets feature high frequency cycles made of two phases: (i) a "make liquidity" phase during which traders post prices (limit orders), and (ii) a "take liquidity" phase during which limit orders are hit by market orders, generating a transaction. The submission of market orders depletes the limit order book of liquidity and ignites a new make/take cycle as it creates transient opportunities for traders submitting limit orders.[1] The speed at which these cycles are completed determines the trading rate, a dimension of market liquidity.

A trader reacts to a transient increase or decline in the liquidity of the limit order book only when she becomes aware of this trading opportunity. Accordingly, the dynamics of trades and quotes in limit order markets is in part determined by traders' monitoring decisions, as emphasized by some empirical studies (e.g., Biais et al. (1995), Sandås (2001) or Hollifield et al. (2004)). For instance, Biais, Hillion, and Spatt (1995) observe that (p.1688): "*Our results are consistent with the presence of limit order traders monitoring the order book, competing to provide liquidity when it is rewarded, and quickly seizing favorable trading opportunities.*"

In practice, traders react with delay (possibly very short) to trading opportunities because monitoring is costly as intermediaries (brokers, market-makers, etc...) have limited monitoring capacity and choose to allocate their limited attention over many markets.[2] Hence, the trading rate and more generally the durations between events (new quotes, trades etc...) are endogenous and determined by a trade-off between the benefit and cost of monitoring. In this paper, we develop a model of trading with imperfect monitoring to study this trade-off and its impact on the trading rate. In the process, we address two sets of related issues.

Firstly, algorithmic trading (the automation of monitoring and orders submission) considerably decreases the cost of monitoring and revolutionizes the way liquidity is provided and consumed. We use our model to study the effects of this evolution on the trading rate, the bid-ask spread, and welfare.

[1] These cycles are studied empirically in Biais, et al. (1995), Coopejans et al.(2003), Degryse et al.(2005), and Large (2007).

[2] For instance, Corwin and Coughenour (2008) show that limited attention by market-makers ("specialists") on the floor of the NYSE affects their liquidity provision.

	Tape A - NYSE Stocks		Tape B - Other Stocks		Tape C - NASDAQ Stocks	
	Make Fee	Take Fee	Make Fee	Take Fee	Make Fee	Take Fee
BATS	-24	25	-24	25	-24	25
EDGX	-25	30	-30	30	-25	30
LavaFlow	-24	27	-24	27	-24	27
Nasdaq	-20	30	-20	30	-20	30
NYSEArca	-23	30	-22	30	-23	30

Table 1: Fees per share (in cents for 100 shares) for limit orders (Make Fee) and market orders (Take Fee) on different trading platforms in the US. A minus sign indicates a rebate. Some platforms use volume-based schedules. The table only shows trading platforms' base pricing. Source: Traders Magazine, August 2009

Secondly, we study the role and effects of so called make/take fees. In each transaction, the "make fee" is the fee charged on the side "making liquidity" (i.e., posting a limit order) and the "take fee" is the fee charged on the side "taking liquidity" (i.e., submitting a market order). For instance, Table 1 gives the make/take fees charged on liquidity makers and liquidity takers for several US equity trading platforms, as of August 2009. All these platforms pay a rebate on executed limit orders and cover this cost by charging a fee on liquidity takers (so called "access fees"). This pricing policy is also used by a few European trading platforms and has been recently adopted in some option markets in the US.

Make/take pricing schedules result in significant monetary transfers between market participants.[3] This practice is very controversial. Market-making firms who use highly automated strategies are generally in favor of make/take pricing, while other market participants have voiced concerns that it could result in excessive fees for liquidity takers.[4] As a result, the SEC decided to cap take fees at \$0.003 per share (30% of the tick size) in equities markets. However, to our knowledge, there is no economic analysis of make/take fees and their effect on market quality. In this paper,

[3] For instance, in each transaction, BATS charges a fee of 0.25 cents per share on market orders and rebates 0.24 cents on executed limit orders (see Table 1). On October 10, 2008, 838,488,549 shares of stocks listed on the NYSE were traded on BATS (about 9% of the trading volume in these stocks on this day); see BATS website: http://www.batstrading.com/. Thus, collectively on this day, limit order traders involved in these transactions collected about \$2.01 million in rebates from BATS while traders submitting market orders paid about \$2.09 million in fees to BATS.

[4] As an example of the controversies raised by these fees, see the petition for rule-making regarding access fees in option markets, addressed by Citadel at the SEC at http://www.sec.gov/rules/petitions/2008/petn4-562.pdf. In this petition, Citadel supports a cap on access fees. For a different viewpoint on make/take fees, see the comments sent by GETCO to the SEC at: http://www.getcollc.com/index.php/getco/commentletters/Schedule of Fees and Charges.pdf.

we fill this gap by providing a theory of make/take fees.

We consider a trading platform with two types of traders: (i) those who post quotes (the "market-makers") and (ii) those who hit these quotes (the "market-takers"). All market participants monitor the market to grab fleeting trading opportunities. Specifically, a market-maker wants to be first to post new quotes after a transient increase in the bid-ask spread and a market-taker wants to be first to hit quotes when the bid-ask spread is tight. In choosing their monitoring intensity, traders on each side trade-off the benefit from a higher likelihood of being first to detect a profit opportunity with the opportunity cost of monitoring. The model has a rich set of testable implications.

First, the speed at which market-makers post good prices and the speed at which these prices are hit are positively related because makers and takers' monitoring decisions reinforce each other. For instance, suppose that an exogenous shock induces market-takers to monitor the market more intensively. Then, market-makers expect more frequent profit opportunities since good prices are hit more quickly. Hence, they have an incentive to monitor more and as a consequence the market features good prices more frequently, which in turn induces market-takers to monitor even more.

We show that this complementarity in traders' speeds of reaction offers a new explanation for the clustering in durations between trades (see for instance Engle and Russell (1998)). It also creates a coordination problem, which results in two equilibria differing in trading activity (i) an equilibrium with no monitoring and no trading and (ii) an equilibrium with monitoring and trading.[5]

Second, the model implies that make/take fees can be used to maximize the trading rate by optimally balancing the speeds of reaction of market-makers and market-takers. To see why, suppose that market-takers' monitoring cost is relatively small while gains from trade when a transaction occurs are equally split between market-makers and market-takers. In this case, in equilibrium, market-takers monitor the market more than market-makers. Thus, good prices take relatively more time to be posted than it takes time for market-takers to hit these prices. The relatively sluggish response of market-makers slows down trading since trades happen when

[5]It is well-known that liquidity externalities create coordination problems among traders, which lead to multiple equilibria with differing levels of liquidity (see Admati and Pfleiderer (1988), Pagano (1989), and Dow (2004) for example). In contrast to the extant literature, our model emphasizes the egg and chicken problem that exists between traders posting quotes on the one hand and traders hitting quotes on the other hand.

the bid-ask spread is tight. To achieve a higher trading rate, the trading platform can reduce its fee on market-makers while increasing its fee on market-takers by the same amount. Its total profit per trade is unchanged but market-makers have now a higher incentive to quickly improve upon unaggressive quotes. Thus, good prices, hence trades, are more frequent.

Following this logic, we find that the optimal make fee relative to the optimal take fee increases in (i) the tick size, (ii) the ratio of the number of market-makers to the number of market-takers, and (iii) the ratio of market-takers' monitoring cost to market-makers' monitoring cost. Indeed, in equilibrium, an increase in these parameters raises the speed at which good prices are posted relative to the speed at which they are hit. Thus, the need to incentivize market-makers is lower.

Third, we study the effect of algorithmic trading by considering the impact of a decrease in monitoring cost in our model. The model implies a strong positive relationship between algorithmic trading and the trading rate (as found empirically in Hendershott, Jones, and Menkveld (2009)). For instance, consider a decrease in the monitoring cost for market-takers. This decrease leads market-takers to hit good prices more quickly, which in itself contributes to increase the trading rate. But, as liquidity is consumed more quickly, market-makers react by supplying liquidity more quickly as well. This feedback effect contributes to increase the trading rate even more.

Algorithmic trading leads to a faster market but its impact on the bid-ask spread is ambiguous. It depends on whether algorithmic trading makes market-takers or market-makers relatively faster. For instance, as just explained, a decrease in market-takers' monitoring cost increases the speed of reaction to changes in the state of the market for both sides. But this increase is stronger for the market-takers. Thus, when market-takers' monitoring cost declines, liquidity is consumed relatively more quickly than it is supplied and as a consequence the bid-ask spread increases on average. In contrast, when market-makers' monitoring cost is reduced, liquidity is supplied relatively more quickly than it is consumed and as a result the bid-ask spread declines on average.

Finally, we find that algorithmic trading is always associated with an increase in welfare measured by the sum of all market participants' expected profits. The reason is that it leads to an increase in the trading rate and therefore the rate at which gains from trade are realized. In contrast, its effect on the expected profit of each

participant depends on whether make/take fees are fixed or not. When make/take fees are fixed, algorithmic trading makes all market participants better off as all traders benefit from an increase in the trading rate. This is not necessarily the case when fees are endogenous. For instance, market-makers are charged a higher fee and market-takers a smaller fee when market-makers' monitoring cost declines. Thus, growth of algorithmic trading on the market-making side makes market-takers better off but, paradoxically, it can make market-makers worse off.

Our study is related to several strands of research. Foucault, Roëll and Sandås (2003) and Liu (2008) provide theoretical and empirical analyses of market-making with costly monitoring. However, the effects in these models are driven by market-makers' exposure to adverse selection. Our paper also contributes to the growing literature on the effects of algorithmic trading (e.g., Biais and Weill (2008), Foucault and Menkveld (2008), Hasbrouck and Saar (2009) or Hendershott, Jones, and Menkveld (2009)). It also relates to the literature on payment for order flow (e.g., Kandel and Marx (1999) or Parlour and Rajan (2003)). But this literature focuses on why rebates for liquidity takers, rather than liquidity makers, can be optimal. In our theory, depending on market structure, either type of rebate can be optimal. Finally, our model contributes to the burgeoning literature on "two-sided markets," i.e., markets in which the volume of transactions depends on the allocation of the total fee per trade between the end-users (see Rochet and Tirole (2006) for a survey).[6]

Section 2 describes the model. In Section 3, we study the equilibrium of the model when make/take fees are fixed. We derive the optimal make/take fees in Section 4 and we provide a detailed discussion of the implications of the model in Section 5. Section 6 concludes. The proofs are in the Appendix.

2 Model

2.1 Market participants

We consider a market for a security with two sides: "market-makers" and "market-takers." Market-makers post quotes (limit orders) whereas market-takers hit these quotes (submit market orders) to complete a transaction. The number of market-makers and market-takers is, respectively, M and N. All participants are risk neutral.

[6] Examples of two-sided markets include videogames platforms, payment card systems etc...See Rochet and Tirole (2006).

We view the market-makers as firms that specialize in high frequency market-making (for instance, Global Electronic Trading company (GETCO), Optiver, Timberhill or Tradebots Systems). The market-taking side represents investors who break their large orders and feed them piecemeal when the bid-ask spread is tight to minimize their trading costs.[7] Both types increasingly use highly automated algorithms to detect and exploit trading opportunities.

In reality, the divide between the market-making side and the market-taking side is not as rigid as assumed here. For instance, electronic market-makers sometimes use market orders. Yet, there is some specialization as electronic market-makers account for a large fraction of liquidity supply on electronic markets.[8] Our assumption captures this feature.[9]

The expected payoff of the security is v_0. Market-takers value the security at $v_0 + L$, where $L > 0$, while market-makers value the security at v_0. Heterogeneity in traders' valuation creates gains from trade (as, for instance, in Duffie et al. (2005) or Hollifield et al. (2004)). As market-takers have a higher valuation, they buy the security from market-makers. Thus, our model describes "the upper half" of the market characterized by limit sell orders and market buy orders. In a more complex model, market-takers could have either high or low valuations relative to market-makers, so that they can be buyers or sellers. This possibility adds some mathematical complexity to the model, but provides no additional economic insight.

Market-makers and market-takers meet on a trading platform with a positive tick-size denoted by Δ, with $\Delta \geq L$. The first price on the grid above v_0 is half a tick above v_0. Let $a \equiv v_0 + \frac{\Delta}{2}$ be this price. All trades take place at this price because market-takers' do not trade at a price higher than a (as, $L \leq \Delta$) and market-makers lose money if they trade at a price smaller than a (as, $a - \Delta < v_0$). Thus, we focus on a "one tick market" as in Parlour (1998) or Large (2009). As in these models, a large number of shares is offered for sale at price $a + \Delta$ by a fringe of competitive traders. The cost of liquidity provision for these traders is higher than

[7] Obizhaeva and Wang (2006) solve the dynamic optimization of such traders, assuming that they exclusively use market orders as we do here.

[8] For instance, Schack and Gawronski (2008) write on page 74 that: "*based on our knowledge of how they do business [...], we believe that they [electronic market-makers] may be generating two-thirds or more of total daily volume today, dwarfing the activity of institutional investors.*"

[9] In some cases, this specialization is imposed by the trading platform. For instance, only designated market-makers can post limit orders on EuroMTS (a trading platform for government bonds in Europe).

for the electronic market-makers and therefore they cannot intervene profitably at price a. Thus, the (half) bid-ask spread is either competitive ($\frac{\Delta}{2}$) or non competitive ($\frac{3\Delta}{2}$), when there is no offer at price a.

There is an upper bound (normalized to one) on the number of shares that can be profitably offered at price a. This upper bound rules out the uninteresting case in which a single market-maker or multiple market-makers offer an infinite quantity at price a. In a more complex model, the upper bound could derive from an upward marginal cost of liquidity provision due, for instance, to exposure to informed trading as in Glosten (1994) or Sandås (2001).

The trading platform charges make/take trading fees each time a trade occurs. The fee (per share) paid by a market-maker is denoted c_m, whereas the fee paid by a market-taker is denoted c_t. We normalize the cost of processing trades for the trading platform to zero so that, per transaction, the platform earns a profit of

$$\overline{c} \equiv c_m + c_t.$$

Introducing an order processing cost per trade is straightforward and does not change the results.

Thus, for each transaction, the gain from trade (L) is split between the parties to the transaction and the trading platform as follows: the market-taker obtains

$$\pi_t = L - \frac{\Delta}{2} - c_t, \tag{1}$$

the market-maker obtains

$$\pi_m = \frac{\Delta}{2} - c_m, \tag{2}$$

and the platform obtains $\bar{c}$. Consequently, the gains from trade accruing to market-makers and market-takers are $\pi_t + \pi_m = L - \bar{c}$. We focus on the case $0 \leq \bar{c} \leq L$ since otherwise at least one side loses money on each trade, and would therefore choose not to participate.

We allow make/take fees to be negative, but such rebates cannot exceed half the tick size. Thus, even in the presence of rebates, it cannot be optimal for market-makers to trade at $a - \Delta$ or for market-takers to trade at $a + \Delta$. As shown in Section 4, in our set-up, this constraint on the size of liquidity rebates is not binding. In practice (see Table 1) rebates are much smaller than half a tick size (a penny in U.S. markets).

At various points in the analysis, we study the effect of a small change in the make/take fees. The existence of a minimum price variation is important for this analysis as it prevents market-makers from neutralizing a change in the fees by adjusting their quotes.[10]

Our setup is clearly very stylized. Yet, it captures in the simplest possible way the essence of the liquidity cycles described in the introduction. Specifically, when there is no quote at a, there is a profit opportunity for market-makers. Indeed, the first market-maker who submits an offer at this price will serve the next buy market order and earns π_m. Conversely, when there is an offer at a, there is a profit opportunity (worth π_t) for a market-taker. After a trade, the bid-ask spread widens. Consequently, the market oscillates between a state in which there is a profit opportunity for market-makers and a state in which there is a profit opportunity for market-takers. Thus, market-makers and market-takers have an incentive to monitor the market. Market-makers are looking for periods when liquidity is scarce and market-takers are looking for periods when liquidity is abundant. Possible extensions of the model are discussed in the conclusion.

2.2 Cycles, Monitoring, and Timing

We now define the notion of "cycles," describe how we model market monitoring, and explain the timing of the game.

Cycles. This is an infinite horizon model with a continuous time line. At each point in time the market can be in one of two states:

1. State E (for Empty)– no offer is posted at a.
2. State F (for Full)– an offer for one share is posted at a.

Thus F is the state in which the (half) bid-ask spread is competitive (the limit order book is Full at price a) whereas E is the state in which the bid-ask spread is not competitive (the limit order book is Empty at price a). The market moves from state F to state E when a market-taker hits the best offer. Thus, the bid-ask spread remains large until a market-maker posts the competitive offer. At this point the bid-ask spread reverts to the competitive level, i.e., the market moves from state E

[10] For instance, if the make fee increases by 1% of the tick size, market-makers cannot neutralize this increase by raising their ask price to $a + 1\% \cdot \Delta$, as this price is not on the grid.

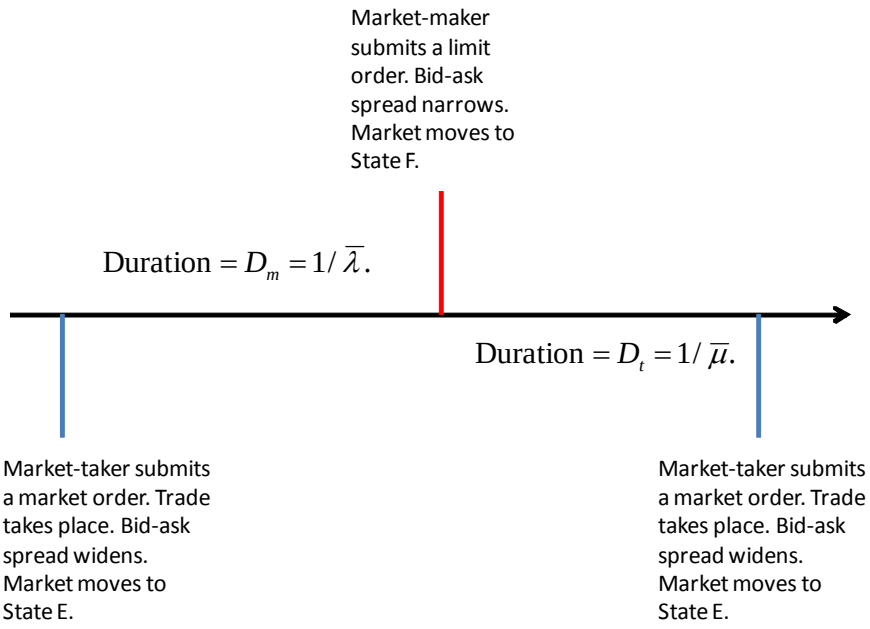

Figure 1: Flow of Events in a Cycle

to state F. Then, the process starts over. We call the flow of events from the moment the market gets into state E until it returns into this state - a "*make/take cycle*" or for brevity just a "cycle." Figure 1 illustrates the flow of events in a cycle.

Monitoring. Each market-maker and each market-taker monitor the market to be the first to detect a profit opportunity on her or his side. We formalize monitoring as follows. Each market-maker $i = 1, ..., M$ inspects the market according to a Poisson process with parameter λ_i, that characterizes her monitoring intensity. As a result, the time between two inspections by market-maker i is distributed exponentially with an average inter-inspection time of $\frac{1}{\lambda_i}$. Similarly, each market-taker $j = 1, ..., N$ inspects the market according to a Poisson process with parameter μ_j.[11] The aggregate monitoring level of the market-making side is

$$\bar{\lambda} \equiv \lambda_1 + ... + \lambda_M,$$

[11] This approach rules out deterministic monitoring such as inspecting the market exactly once every certain number of minutes. In reality, many unforeseen events can capture the attention of a market-maker or a market-taker, be it human or a machine. For humans, the need to monitor several securities as well as perform other tasks precludes evenly spaced inspections. Computers face similar constraints as periods of high transaction volume, and unexpectedly high traffic on communication lines prevent monitoring at exact points in time.

and the aggregate monitoring level of the market-taking side is

$$\bar{\mu} \equiv \mu_1 + ... + \mu_N.$$

When a market-maker inspects the market she learns whether it is in state E or F. If the bid-ask spread is not competitive (state E) then she posts an offer at a. If it is competitive (state F), the market-maker stays put until her next inspection. Similarly, a market-taker submits a market order when, upon inspection, he observes that the bid-ask spread is competitive, and stays put until the next inspection otherwise.[12]

Thus, the duration from a trade to a competitive quote (state E to state F) is exponentially distributed with parameter $\overline{\lambda}$ and the duration from a competitive quote to a trade (state F to state E) is exponentially distributed with parameter $\overline{\mu}$. As traders' monitoring levels are endogenous (see below), the distributions of inter-event durations are endogenous in the model. We denote by $\mathcal{D}_m \equiv \frac{1}{\bar{\lambda}}$, the expected duration from the time the bid-ask spread widens (state E) until it becomes competitive (state F) and by $\mathcal{D}_t \equiv \frac{1}{\bar{\mu}}$, the expected duration from the time the bid-ask spread becomes small until a trade takes place. Notice that $\mathcal{D}_m$ is a measure of market resiliency since it is low when an increase in the bid-ask spread after a trade is quickly corrected with a more competitive offer.[13]

Thus, on average, the duration between two trades (the average duration of a cycle) is

$$\mathcal{D}\left(\bar{\lambda}, \bar{\mu}\right) \equiv \mathcal{D}_m + \mathcal{D}_t = \frac{1}{\bar{\lambda}} + \frac{1}{\bar{\mu}} = \frac{\bar{\lambda} + \bar{\mu}}{\bar{\lambda} \cdot \bar{\mu}}, \tag{3}$$

and the trading rate, i.e., the average number of transactions per unit of time, is

$$\mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right) \equiv \frac{1}{\mathcal{D}\left(\bar{\lambda}, \bar{\mu}\right)} = \frac{\bar{\lambda} \cdot \bar{\mu}}{\bar{\lambda} + \bar{\mu}}. \tag{4}$$

The trading rate depends on traders' aggregate monitoring levels and increases when either $\bar{\lambda}$ or $\bar{\mu}$ increase.

In practice, monitoring can be manual, by looking at a computer screen, or automated by using automated algorithms. For humans, the need to monitor several stocks contemporaneously constrains the amount of attention dedicated to a specific

[12] Hall and Hautsch (2007) model the arrival of buy and sell market orders as a Poisson Process with state-dependent intensities. They find empirically that these intensities are higher when the bid-ask spread is tight. This empirical finding is consistent with our assumption that market takers submit their market orders when the bid-ask spread is competitive.

[13] See, for instance, Foucault et al.(2005), Large (2007) and Roşu (2008) for analysises of market resiliency.

stock. Computers also have fixed processing power that must be allocated over potentially hundreds of stocks and millions of pieces of information. Prioritization of this process is conceptually similar to the allocation of attention across different stocks by a human market-marker. Hence, in all cases, monitoring one market is costly, because it reduces the monitoring capacity (or processing power) available for other markets.

Thus, we assume that, over a time interval of length T, a market-maker choosing a monitoring intensity λ_i bears a monitoring cost:

$$C_m(\lambda_i) \equiv \frac{1}{2}\beta\lambda_i^2 T \quad \text{for } i = 1, ..., M. \tag{5}$$

Similarly, the cost of monitoring for market-taker j over an interval of time of length T is:

$$C_t(\mu_j) \equiv \frac{1}{2}\gamma\mu_j^2 T \quad \text{for } j = 1, ..., N. \tag{6}$$

We say that market-makers' (resp. market-takers') monitoring cost becomes lower when β (resp. γ) decreases. This decline can be a result, for example, of automation in the monitoring process and technological improvements that allow much faster access to information. Thus, below, we analyze the effect of algorithmic trading on the trading process by considering the effect of a reduction in β and γ. Parameters γ and β must remain strictly positive, but all our results hold even if γ and β become infinitesimal so that the cost of monitoring appears negligible (which maybe is the case with algorithmic trading). In fact, what matters for most of our implications (e.g., the relative sizes of make/take fees) is not the absolute size of γ and β, but their relative size, $\frac{\gamma}{\beta}$. We denote this ratio by $r \equiv \frac{\gamma}{\beta}$.

Timing. In reality, traders can change their monitoring intensities as market conditions change, whereas trading fees are fixed in the short-run. For this reason, we assume that traders choose their monitoring intensities after observing the fees set by the trading platform. Hence the trading game unfolds in three stages as follows:

Stage 1: The trading platform chooses its make/take fees c_m and c_t.

Stage 2: Market-makers and market-takers simultaneously choose their monitoring intensities λ_i and μ_j.

Stage 3: From this point onward, the game is played on a continuous time line indefinitely, with the monitoring intensities and fees determined in Stages 1 and 2.

2.3 Objective functions and equilibrium

We now describe market participants' objective functions and define the notion of equilibrium that we use to solve for players' optimal actions in each stage.

Objective functions. Recall that a make/take cycle is a flow of events from the time the market is in state E until it goes back to this state. Each time a make/take cycle is completed a transaction occurs. The probability that market-maker i is active in this transaction is the probability that she is first to post a competitive offer at price a after the market entered in state E. Given our assumptions on the monitoring process, this probability is $\frac{\lambda_i}{\lambda_1+...+\lambda_M} = \frac{\lambda_i}{\bar{\lambda}}$. Thus, in each cycle, the expected profit (gross of monitoring costs) for market-maker i is $\frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m = \frac{\lambda_i}{\bar{\lambda}} \left(\frac{\Delta}{2} - c_m\right)$.

Let $\tilde{n}_T$ be the (random) number of completed transactions (cycles) until time T. The expected payoff to market-maker i until time T (net of monitoring costs) is

$$\Pi_i(T) = E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} \frac{\lambda_i}{\bar{\lambda}} \pi_m) - \frac{1}{2}\beta\lambda_i^2 T.$$

As is common in infinite horizon Markovian models, we assume that each player maximizes his/her long-term (steady-state) payoff per unit of time. Thus, market-maker i chooses his monitoring intensity to maximize

$$\Pi_{im} \equiv \lim_{T\to\infty} \frac{\Pi_i(T)}{T} = \lim_{T\to\infty} \frac{E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} p_i \pi_m)}{T} - \frac{1}{2}\beta\lambda_i^2. \tag{7}$$

Recall that $\mathcal{D}\left(\bar{\lambda}, \bar{\mu}\right)$ is the expected duration of a cycle. A standard theorem from the theory of stochastic processes (often referred to as the "Renewal Reward Theorem" see Ross (1996), p. 133) implies that

$$\lim_{T\to\infty} \frac{E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} p_i \pi_m)}{T} = \frac{\frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m}{\mathcal{D}\left(\bar{\lambda}, \bar{\mu}\right)} = \frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right).$$

Thus, the objective function of market-maker i (equation (7)) is:

$$\Pi_{im} = \frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right) - \frac{1}{2}\beta\lambda_i^2. \tag{8}$$

This is intuitive: the expected profit of a market-maker per unit of time is his expected profit per transaction ($\frac{\lambda_i}{\bar{\lambda}} \cdot \pi_m$) times the trading rate, less the monitoring cost. In a similar way, the objective function of market-taker j can be written as

$$\Pi_{jt} = \frac{\mu_j}{\bar{\mu}} \cdot \pi_t \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right) - \frac{1}{2}\beta\mu_j^2. \tag{9}$$

Finally, in each cycle, the trading platform earns a fee $\bar{c}$. Thus, similar arguments show that the objective function of the exchange is:

$$\Pi_e \equiv \bar{c} \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right) = (c_m + c_t) \cdot \mathcal{R}\left(\bar{\lambda}, \bar{\mu}\right). \tag{10}$$

Liquidity Externalities and Cross-Side Complementarities. An increase in the aggregate monitoring level of one side exerts a positive externality on the other side since $\frac{\partial \Pi_{im}}{\partial \bar{\mu}} > 0$ and $\frac{\partial \Pi_{jt}}{\partial \bar{\lambda}} > 0$. Intuitively, a higher aggregate monitoring intensity for market-makers (resp., market-takers) increases the rate at which market-takers (resp., market-makers) find trading opportunities and therefore make the latter better-off. Moreover, the marginal benefit of monitoring for traders on one side increases in the aggregate monitoring level of traders on the other side since $\frac{\partial^2 \Pi_{im}}{\partial \bar{\mu} \partial \lambda_i} > 0$ and $\frac{\partial^2 \Pi_{jt}}{\partial \bar{\lambda} \partial \mu_j} > 0$. For this reason, market-makers (resp., market-takers) will check the state of the market more frequently when they expect market-takers (resp. market-makers) to check the state of the market more frequently. Thus, market-makers and market-takers' monitoring decisions are self-reinforcing. In other words, liquidity supply begets liquidity demand and vice versa. As we shall see, this "cross-side complementarity" has important implications.

In contrast, an increase in one trader's monitoring level hurts the traders who are on his or her side. That is, $\frac{\partial \Pi_{im}}{\partial \lambda_j} < 0$ and $\frac{\partial \Pi_{it}}{\partial \mu_j} < 0$ (for $j \neq i$). This effect captures the fact that traders on the same side are engaged in a "horse race" to be first to detect a trading opportunity. In reality, keeping up with this race is a key reason for automating order submission.[14]

Equilibrium. The strategies for the market-makers and market-takers are their monitoring intensities λ_i and μ_j respectively. A strategy for the trading platform is a menu of make/take fees (c_m, c_t). We solve the model backwards. First, for given fees, we look for Nash equilibria in monitoring intensities in Stage 2. A Nash equilibrium

[14] See for instance "Tackling latency-the algorithmic arms race," IBM Global Business Services report.

in this stage is a vector of monitoring intensities $(\lambda_1^*, \ldots, \lambda_M^*, \mu_1^*, \ldots, \mu_N^*)$ such that for all $i = 1, \ldots, M$, λ_i^* maximizes market-maker i's expected profit per unit of time (given by (8)), and for all $j = 1, ..., N$, μ_j^* maximizes market-taker j's expected profit per unit of time (given by (9)), taking the monitoring intensities of all other traders as given. Second, given a Nash equilibrium in the monitoring intensities, we solve for the make/take fees (c_m^*, c_t^*) that maximize the trading platform's expected profit (equation (10)).[15]

3 Equilibria with Fixed Fees

In this section we study the equilibrium monitoring intensities for given fees (c_m, c_t). For all parameters values, the model has two equilibria: (i) an equilibrium with no trade and (ii) an equilibrium with trade. Indeed, the complementarity in monitoring decisions between market-makers and market-takers leads to a coordination problem.

To see this point, consider how the no-trade equilibrium arises. If market-makers believe that market-takers will not monitor the trading platform, then they optimally choose not to monitor the platform as well ($\lambda_i^* = 0$) since monitoring is costly. Symmetrically, if market-takers expect market-makers to pay no attention to the trading platform then they optimally choose to be inactive ($\mu_j^* = 0$). Thus, traders' beliefs that the other side will not be active are self-fulfilling, which result in a no-monitoring, no-trade equilibrium.

Proposition 1 *For all parameters, there is an equilibrium in which traders do not monitor:* $\lambda_i^* = \mu_j^* = 0$ *for all* $i \in \{1, \ldots, M\}$ *and* $j \in \{1, \ldots, N\}$. *The trading rate in this equilibrium is zero.*

The second equilibrium does involve monitoring and trade. To describe this equilibrium, let

$$z \equiv \frac{\pi_m}{\pi_t}\frac{\gamma}{\beta} = \frac{\pi_m}{\pi_t} \cdot r. \tag{11}$$

When $z > 1$ (resp. $z < 1$), the ratio of profits to costs per cycle is larger for market-makers (resp. market-takers).

[15] The fees c_m and c_t affect traders' objective functions (equations (8) and (9)) directly (through their effect on π_m and π_t) and indirectly, through their effect on traders' monitoring levels in equilibrium.

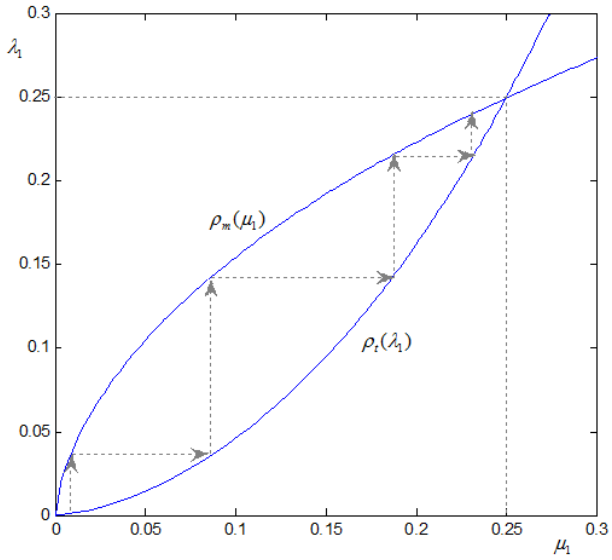

Figure 2: Low and High Liquidity Regimes. The market-maker's best-response function, $\rho_m(\cdot)$, is plotted as a function of μ_1 (the horizontal axis), whereas the market-taker's best-response function, $\rho_t(\cdot)$, is plotted as a function of λ_1 (the vertical axis).

Proposition 2 *For all parameters, there exists a unique equilibrium with trade. In this equilibrium, traders' monitoring intensities are given by*

$$\lambda_i^* = \frac{M + (M-1)\,\Omega^*}{(1+\Omega^*)^2} \cdot \frac{\pi_m}{M\beta} \qquad i = 1, \ldots, M \tag{12}$$

$$\mu_j^* = \frac{\Omega^* \left((1+\Omega^*)\,N - 1\right)}{(1+\Omega^*)^2} \cdot \frac{\pi_t}{N\gamma} \qquad j = 1, \ldots, N \tag{13}$$

where Ω^* is the unique positive solution to the cubic equation

$$\Omega^3 N + (N-1)\Omega^2 - (M-1)\,z\Omega - Mz = 0. \tag{14}$$

Moreover, in equilibrium, $\frac{\bar{\lambda}^*}{\bar{\mu}^*} = \Omega^*$.

As an illustration, consider the case $M = N = 1$. Figure 2 plots the best-response functions, denoted by $\rho_m(\mu_1)$ and $\rho_t(\lambda_1)$, for the market-maker and market-taker respectively, when $\pi_m = \pi_t = 0.5$, $\beta = \gamma = 0.5$.[16] The two best-response functions

[16] That is, $\rho_m(\mu_1)$ (resp. $\rho_t(\lambda_1)$) gives the optimal monitoring level of the market-maker (resp. market-taker) given that the market-taker's monitoring level is μ_1 (resp. λ_1).

meet at two points (0,0) and (0.25,0.25), which are the two equilibria in this example (from Propositions 1 and 2). It can be verified that the slope of both reaction functions at zero is infinite. Thus, in this situation, an infinitesimal increase in, say, the market-taker's monitoring, μ_1, triggers a relatively large increase in the market-maker's monitoring, λ_1, which in turn leads to a larger μ_1, and so on. Along this path, the trading rate on the platform builds up since it increases with monitoring levels on either side. This process ends when traders' monitoring levels converge to their level in Proposition 2. These dynamics could be used to interpret the evolution of trading on new trading platforms (e.g., Chi-X or Turquoise in Europe). The complementarity between market-makers and market-takers is key here since it explains the evolution from a low (zero) to a high trading rate equilibrium. Thus, measuring empirically this complementarity is important. We come back on this question in Section 5.

As the no-trade equilibrium us unstable, in the rest of the paper we focus on the properties and implications of the equilibrium with trade.

Corollary 1 *In the unique equilibrium with trade,*

1. *The aggregate monitoring level of each side increases in the number of participants on either side* ($\frac{\partial \bar{\lambda}^*}{\partial N} > 0$, $\frac{\partial \bar{\lambda}^*}{\partial M} > 0$, $\frac{\partial \bar{\mu}^*}{\partial N} > 0$, $\frac{\partial \bar{\mu}^*}{\partial M} > 0$) *and decreases in (i) monitoring costs* ($\frac{\partial \bar{\lambda}^*}{\partial \beta} < 0$, $\frac{\partial \bar{\lambda}^*}{\partial \gamma} < 0$, $\frac{\partial \bar{\mu}^*}{\partial \beta} < 0$, $\frac{\partial \bar{\mu}^*}{\partial \gamma} < 0$) *or (ii) the fee per trade charged on either side* ($\frac{\partial \bar{\lambda}^*}{\partial c_m} < 0$, $\frac{\partial \bar{\lambda}^*}{\partial c_t} < 0$, $\frac{\partial \bar{\mu}^*}{\partial c_m} < 0$, $\frac{\partial \bar{\mu}^*}{\partial c_t} < 0$).

2. *The trading rate decreases in (i) the monitoring costs* ($\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial \beta} < 0$ *and* $\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial \gamma} < 0$) *or the trading fees* ($\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m} < 0$ *and* $\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_t} < 0$) *and (ii) increases in the number of participants on either side* ($\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial M} > 0$ *and* $\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial N} > 0$).

To understand the first part of the corollary, consider an increase in the monitoring cost for market-makers. This increase reduces their individual monitoring levels, other things equal. Consequently, the marginal benefit of monitoring for market-takers is smaller as they are less likely to find a good price when they inspect the market. Thus, market-takers monitor the market less intensively, even though their own monitoring cost has not changed, which in turn induces market-makers to monitor even less. The same reasoning applies for an increase in the trading fee or the

number of participants on one side.[17]

Corollary 1 is interesting for two reasons. First, it implies a positive correlation between inter-event durations since these durations are inversely related to aggregate monitoring levels ($\mathcal{D}_m = \frac{1}{\bar{\lambda}}$ and $\mathcal{D}_t = \frac{1}{\bar{\mu}}$). For instance, consider a positive shock to the number of market-takers. This shock increases traders' aggregate monitoring on both sides of the market and eventually reduces all inter-event durations. Hence, the complementarity between liquidity suppliers and liquidity demanders offers a possible explanation for the clustering in durations observed in high frequency data (more on this in Section 5.1).

Second, Corollary 1 implies that the effect of a shock to the parameters of one side (e.g., c_m or β) on the trading rate is magnified by the reaction of the other side. For instance, if market-makers monitoring cost is reduced (β decreases), then the reaction time of market-makers is reduced and the trading rate increases. But precisely for this reason, market-takers monitor more and their reaction time is reduced as well, which increases the trading rate even more. As explained in Section 5.4, this amplification effect implies that the development of algorithmic trading should have a first order effect on the trading rate.

Corollary 2 *In equilibrium, for fixed fees, the market-making side monitors the market more (less) intensively than the market-taking side ($\bar{\lambda}^* > \bar{\mu}^*$) if and only if $\frac{z(2M-1)}{2N-1} > 1$. If $\frac{z(2M-1)}{2N-1} = 1$, the market-making and the market-taking sides have identical monitoring intensities.*

We define the *velocity ratio* as the relative speed of reaction of the market-making side vs. the market-taking side: $\mathcal{V} \equiv \frac{\mathcal{D}_t}{\mathcal{D}_m} = \frac{\bar{\lambda}^*}{\bar{\mu}^*}$. From Corollary 2, if $\frac{z(2M-1)}{(2N-1)} > 1$, then market-makers post good prices after a trade more quickly than market-takers hit these prices and $\mathcal{V} > 1$. For instance, if $M = N$ and $\frac{\pi_m}{\beta} > \frac{\pi_t}{\gamma}$, the market-making side reacts more quickly than the market-taking side because market-makers' cost of missing a trading opportunity is relatively higher.

[17] When the number of participants on one side increases, the individual monitoring levels of the market participants on this side may decrease. Indeed, as more traders on one side compete for trading opportunities, the likelihood of being first to grab this opportunity declines. This effect decreases the incentive to monitor of each participant on the side that becomes thicker. Yet, it is small as it is offset by the cross-side complementarity effect, which is conducive to more monitoring by each participant. Thus, when the number of participants on one side increases, the aggregate monitoring of both sides increases.

As shown in the next section, this observation is important to understand the optimal pricing policy for the trading platform. Intuitively, a situation in which the velocity ratio is too large or too small is suboptimal for the platform. Indeed, it means that one side is very quick in taking advantage of trading opportunities but this velocity does not translate into a high trading rate if the other side is relatively very slow. Thus, in this situation, it is optimal for the trading platform to alter its fees so as to reduce the imbalance between the velocities of both sides. The next corollary shows how the velocity ratio changes when trading fees or other parameters change.

Corollary 3 *The velocity ratio is (a) positively related to γ, c_t and M, and (b) negatively related to β, c_m and N.*

Thus, the trading platform through the choice of its make-take fees determines both the trading rate (Corollary 1) and the velocity ratio (Corollary 3). For instance, the trading platform can reduce the velocity ratio without changing its revenue per trade by increasing the make fee while reducing the take fee.

The thick market case. In general we do not have an explicit solution for traders' monitoring levels because we cannot solve for Ω^* in closed-form (Ω^* is the unique positive root of equation (14)). However, there are a few cases in which a closed form solution can be obtained (e.g., $M = N = 1$).[18] One interesting and useful case is when the number of participants on both sides becomes very large (both M and N tend to infinity) but the size of the market-making side *relative* to the size of the market-taking side, $q \equiv \frac{M}{N}$, remains fixed. We refer to this case as "the thick market case." In this case, we deduce from equation (14) that[19]

$$\Omega^{\infty} \equiv \lim_{M \to \infty} \Omega^* = (zq)^{\frac{1}{2}} . \tag{15}$$

Using this observation and Proposition 2, the next corollary provides closed-form expressions for traders' monitoring levels when the market is thick on both sides.

[18] Using Proposition 2, it is easy to verify that when $M = N = 1$, $\Omega^* = z^{\frac{1}{3}}$. Thus, $\lambda_1^* = \left(1 + z^{\frac{1}{3}}\right)^{-2} \cdot \left(\frac{\pi_m}{\beta}\right)$ and $\mu_1^* = \left(1 + z^{-\frac{1}{3}}\right)^{-2} \cdot \left(\frac{\pi_t}{\gamma}\right)$.

[19] To see this, note that equation (14) implies $z = \frac{\Omega^{*3}\frac{M}{q} + (\frac{M}{q} - 1)\Omega^{*2}}{(M-1)\Omega^* + M}$. Equation (15) follows by taking the limit as $M \to \infty$.

Corollary 4 *(monitoring levels in the thick market case) Let $q > 0$ be fixed, and assume $N = \frac{M}{q}$. Then,*

$$\begin{aligned} \lambda_i^{\infty} &\equiv \lim_{M\to\infty} \lambda_i^* = \frac{1}{1+(zq)^{\frac{1}{2}}} \cdot \frac{\pi_m}{\beta} \quad i = 1,2,3,... \\ \mu_j^{\infty} &\equiv \lim_{M\to\infty} \mu_j^* = \frac{1}{1+(zq)^{-\frac{1}{2}}} \cdot \frac{\pi_t}{\gamma} \qquad j = 1,2,3,... \end{aligned} \tag{16}$$

It is worth stressing that traders' monitoring levels when the number of market participants is finite quickly converge to their limit levels when the market is thick.[20] Hence, monitoring levels in the thick market case can be used to obtain good approximations of the market behavior even for relatively low values of M and N.

4 Determinants of the Make/Take Fees

Now, we study the fees set by the trading platform. In most of the analysis, we fix the total fee, $\bar{c}$, as we are mainly interested in the optimal breakdown of this fee between makers and takers. We refer to $c_m - c_t$ as the *make/take spread.* The make/take spread is positive (negative) if the market-making side pays a higher (lower) fee than the market-taking side. Our goal is to understand how the exogenous parameters of the model (the tick size, the monitoring costs, and the relative number of participants on each side) affect the make/take fees. For instance, we study the conditions under which the optimal make-take spread is negative ($c_m < c_t$), as often observed in reality (see Table 1 in the introduction).

As explained in Section 2.3, for a given total fee $\bar{c}$, the objective function of the trading platform is to set make/take fees (c_m^*, c_t^*) that solve,

$$\begin{aligned} \max_{c_m,c_t} \ \Pi_e &= (c_m + c_t)\, \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*) \\ s.t: &\quad c_m + c_t = \bar{c}, \end{aligned} \tag{17}$$

where $\bar{\lambda}^* = M\lambda_i^*$, $\bar{\mu}^* = N\mu_j$, and λ_i^* and μ_j^* are given by Proposition 2. Trading fees affect traders' monitoring decisions and thereby the trading rate (Corollary 1). The first order conditions for the trading platform's optimization problem impose that

$$\frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_m} = \frac{\partial \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{\partial c_t}. \tag{18}$$

[20] Formally, we can show that $\Omega^{\infty} - \Omega^*$ is $O\left(\frac{1}{M}\right)$, which means that the error in using Ω^{∞} to approximate for Ω^* is on the order of magnitude of $\frac{1}{M}$. The proof is available upon request.

That is, the trading platform chooses its fee structure so as to equalize the marginal (negative) impact of an increase in each fee on the trading rate.

We denote by η_{mm} and η_{mt} the elasticities of the aggregate monitoring levels of the market-taking side and the market-making side to the fee charged on *market-makers*. Similarly, η_{tt} and η_{tm} are the elasticities of traders' aggregate monitoring level to the take fee. Formally

$$\begin{aligned} \eta_{mm} &\equiv \left(\frac{\partial \log(\bar{\lambda}^*)}{\partial c_m}\right) c_m \quad \text{and} \quad \eta_{mt} \equiv \left(\frac{\partial \log(\bar{\lambda}^*)}{\partial c_t}\right) c_t, \\ \eta_{tt} &\equiv \left(\frac{\partial \log(\bar{\mu}^*)}{\partial c_t}\right) c_t \text{ and } \eta_{tm} \equiv \left(\frac{\partial \log(\bar{\mu}^*)}{\partial c_m}\right) c_m. \end{aligned} \tag{19}$$

The cross-side complementarity implies that $\eta_{mt} < 0$ and $\eta_{tm} < 0$ (when fees are positive). Using equation (18), we obtain the following result.

Proposition 3 *For each level $\bar{c}$ of the total fee charged by the platform, the optimal make/take fees satisfy:*

$$\begin{aligned} c_m^* &= \left(\frac{h}{h+1}\right)\bar{c}, \\ c_t^* &= \bar{c} - c_m^* = \left(\frac{1}{h+1}\right)\bar{c}, \end{aligned} \tag{20}$$

where $h \equiv \frac{(\bar{\lambda}^*)^{-1}\eta_{mm} + (\bar{\mu}^*)^{-1}\eta_{tm}}{(\bar{\lambda}^*)^{-1}\eta_{mt} + (\bar{\mu}^*)^{-1}\eta_{tt}}$.

Thus, it is optimal to charge different fees on market-makers and market-takers, unless $h = 1$. Interestingly, the optimal fee structure (h) depends on the cross-side elasticities of the aggregate monitoring levels to the fees. Indeed, an increase in the fee on one side has a "double" negative effect on the trading rate since it also, indirectly, negatively affects monitoring of the other side.

Proposition 3 does not provide a closed-form solution for the trading fees since the elasticities of monitoring levels to trading fees depend on the fees. However, we can obtain an analytical solution when the market is thick. Therefore, we first study the effects of the exogenous parameters on the make-take fees in this case (Section 4.1). Then, using numerical simulations, we show that the conclusions obtained in this polar case are robust for arbitrary values of M and N (Section 4.2).

4.1 Optimal fees when the market is thick

As shown in Corollary 4, traders' individual monitoring levels remain strictly positive when the market is thick (i.e., M and N tend to infinity, but $\frac{M}{N} = q$). Thus, as

the market becomes thick, traders' aggregate monitoring levels and the trading rate explode. Yet, the fee structure that maximizes the trading rate converges to a well-defined limit, as shown in our next proposition.

Proposition 4 *In the thick market case, the trading platform optimally allocates its fee $\bar{c}$ between the market-making side and the market-taking side as follows:*

$$c_m^* = \frac{1}{2}\left(\Delta - \frac{2(L-\bar{c})}{(1+(qr)^{\frac{1}{3}})}\right) \quad \textit{and } c_t^* = \bar{c} - c_m^*. \tag{21}$$

For these fees,

$$\pi_m^* = \frac{L-\bar{c}}{(1+(qr)^{\frac{1}{3}})} \quad \textit{and} \quad \pi_t^* = \frac{L-\bar{c}}{(1+(qr)^{-\frac{1}{3}})}, \tag{22}$$

and the equilibrium monitoring intensities are:

$$\lambda_i^\infty = \frac{L-\bar{c}}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^2} \quad \textit{and} \quad \mu_j^\infty = \frac{L-\bar{c}}{\gamma\left(1+(qr)^{-\frac{1}{3}}\right)^2} \quad \textit{for } i,j = 1,2,... \tag{23}$$

Using these results we can study how the tick size, the monitoring costs and the ratio of market participants on both sides determine the optimal make/take spread $(c_m^* - c_t^*)$. Let $\bar{\Delta}(q,r) \equiv 2(L-\bar{c})\left(1+(qr)^{\frac{1}{3}}\right)^{-1} + \bar{c}$. Using equation (21), we obtain the following implication.

Corollary 5 *In the thick market case, the make-take spread increases with (i) the tick size, Δ; (ii) the relative size of the market-making side, q; and (iii) the relative monitoring cost for the market-taking side, r. Moreover the make-take spread is negative if and only if $\Delta < \bar{\Delta}(q,r)$.*

Figure 3 illustrates the set of parameters for which the make/take spread is negative or positive. The make-take spread is more likely to be negative when (i) the tick size is small, (ii) the number of market-makers is relatively small, or (iii) the monitoring cost for market-makers is relatively large. Hence, the optimal pricing policy follows a simple principle: the make fee should increase relative to the take fee when a change in parameters raises market-makers' aggregate monitoring. In other words, the trading platform uses its make/take fees to correct too large imbalances in the velocities of the market-making side and the market-taking side.

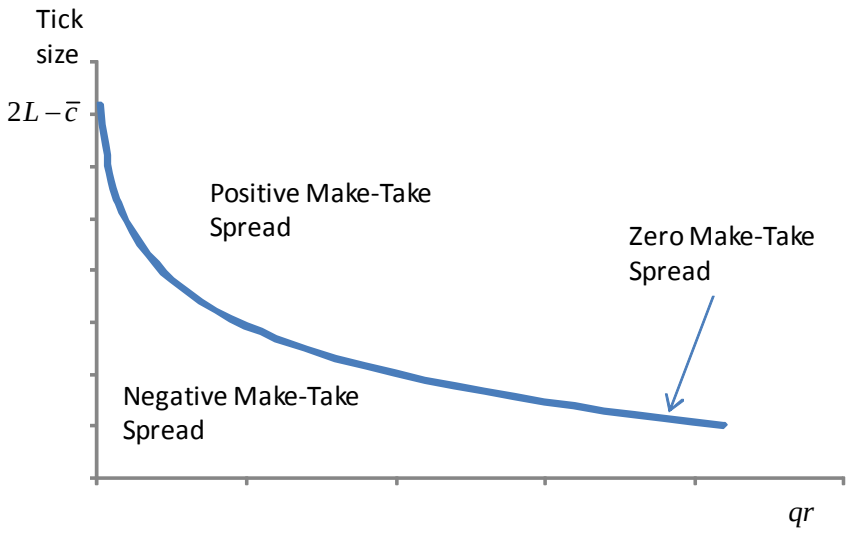

Figure 3: Determinants of the relative sizes of the make/take fees

The reason for this is intuitive. For instance, suppose that the tick size is large. In this case, market-makers obtain a relatively high fraction of the gains from trade when they participate in a transaction (other things equal, $\frac{\pi_m}{\pi_m+\pi_t}$ increases in Δ). Thus, market-makers have a high incentive to monitor the market while market-takers have a relatively low incentive to do so.[21] As a result, good prices are posted very quickly after a trade but market-takers are relatively slow to hit these prices. In this case, it is optimal for the platform to lower its fee on market-takers, to accelerate their response to good prices, and increase its fee on market-makers since their corrective response to wide spreads is already quick. Thus, its revenue per trade is unchanged but the trading rate is higher and finally the expected profit of the platform per unit of time is higher.

Hence, the platform responds to an increase in the tick size (which fosters more monitoring by market-makers, other things equal) by raising its make fee and cutting its take fee. As a result, the make-take spread increases. The effect of other parameters (q or $r = \frac{\gamma}{\beta}$) on the make/take spread can be interpreted in the same way. Intuitively, an increase in the relative size of the market-making side (a higher q) or a decrease in its relative monitoring cost (a higher r) result in a higher monitoring

[21] Other authors have noticed that a large tick size encourage liquidity providers to be active in a stock. See, for instance, Harris (1990), Angel (1997) and Easley, O'Hara and Saar (2001) for an empirical test.

intensity for market-makers relative to market-takers, other things equal. Thus, to balance the speed of reactions of both sides, it is optimal for the the trading platform to raise its fee on the market-making side when q or r increase.

Equation (21) implies that market-makers (resp. market-takers) are optimally subsidized (they pay a negative trading fee) when the tick size is small (resp. large) enough. However, if one side is subsidized, the optimal rebate is always strictly larger than half the tick size ($\text{Min}\{c_m^*, c_t^*\} > -\frac{\Delta}{2}$ since $\overline{c} \leq L \leq \Delta$). Thus, the constraint we imposed on the size of these rebates is not binding for the platform, as mentioned earlier (see Section 2.1).

Now consider the optimal choice of the total fee, $\bar{c}$. Clearly, the platform's optimization problem can be decomposed into two steps: (i) choose the optimal make/take fees for a given $\bar{c}$ (we solved this problem) and (ii) choose the optimal $\bar{c}$. Observe that the optimal make/take fees, (c_m^*, c_t^*), increase in $\bar{c}$, and recall that the trading rate decreases in both the make fee and the take fee (Corollary 1). Thus, in the second step, the trading platform faces the standard price-quantity trade-off: by raising $\bar{c}$, the trading platform gets a larger revenue per trade but it decreases the rate at which trades occur. The next corollary provides the optimal value of $\bar{c}$ for the trading platform.

Corollary 6 *In the thick market case, the trading platform maximizes its expected profit by setting its total trading fee at $\bar{c} = L/2$ and by splitting this fee between both sides as described in Proposition 4.*

In contrast to the make/take fees, the optimal total fee for the platform is independent of the tick size, traders' relative monitoring costs and the relative size of the market-making side. Thus, our results regarding the effect of Δ, q, and r hold even if $\bar{c}$ is optimally set by the trading platform. Finally, note that our findings regarding the optimal make/take fees hold for any level of the total fee, $\bar{c}$. Thus, they would hold as well if this fee were arbitrarily capped at some level (even very low).

4.2 Optimal Fees: General Case

As explained previously, we cannot obtain an analytical solution for the optimal make/take fees for arbitrary values of M and N. However, we can numerically solve for these fees, using the characterization of traders' monitoring levels in Proposition 2. Using this approach, we have checked through extensive numerical simulations

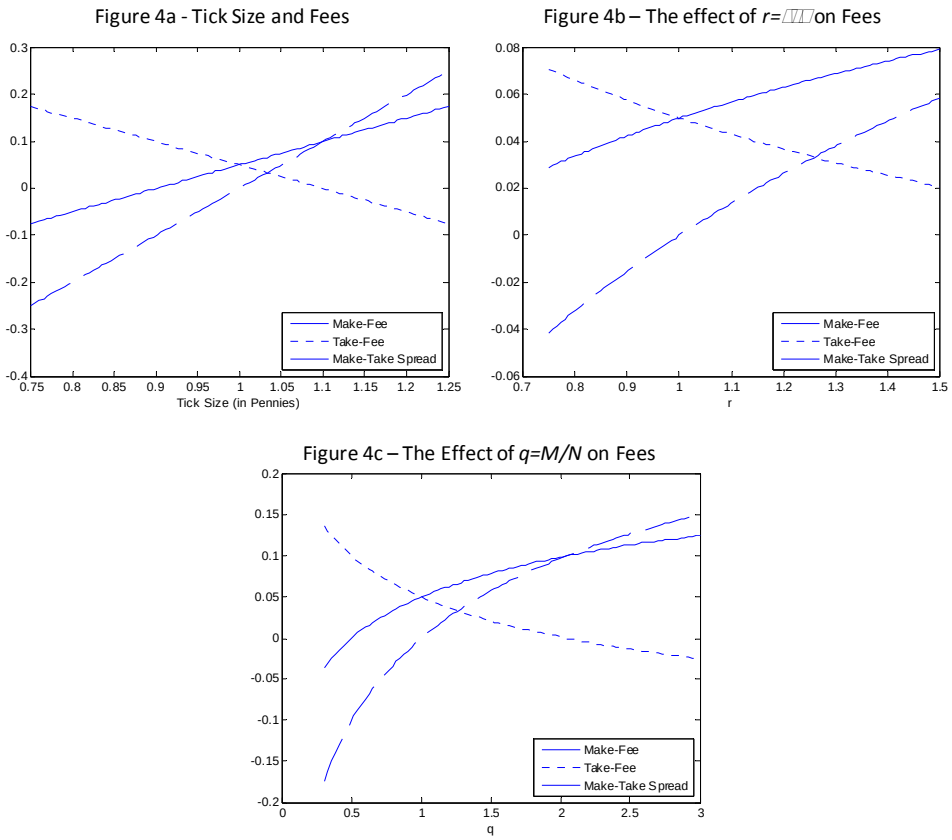

Figure 4: Determinants of the Make/Take Fees

that the comparative static results obtained in the thick market case are robust.[22] As an illustration, consider the simulations reported in Figure 4 (where the baseline values of the parameters are: $M = N = 10$; $\gamma = \beta = 1$; $\Delta = 1$ (1 penny); $L = 1$ (1 penny); $\bar{c} = 0.1$ (0.1 pennies).

In each panel, we plot the take fee (dotted line), the make fee (plain line) and the make-take spread (dashed line) as a function of the tick size, Δ (Figure 4a), the ratio of monitoring costs, $r = \frac{\gamma}{\beta}$ (Figure 4b) and the ratio of the number of market-makers to the number of market-takers, $q = M/N$ (Figure 4c). Clearly the effects of these variables on the optimal make/take fees and the make-take spread are as described in Corollary 5. Hence, the conclusions of this corollary appear robust for arbitrary

[22] The case in which $M = N = 1$ is another case in which we can solve for the optimal make/take fees. The expressions for the fees in this case are very similar to those obtained in Proposition 4. The only difference is that (i) $q = 1$ and (ii) (qr) appears with the exponent $\left(\frac{1}{4}\right)$ instead of $\left(\frac{1}{3}\right)$ in all formulae.

values of the number of participants on either side.

5 Empirical Implications

5.1 Inter-Event Durations and Clustering

It is well-known that long (short) durations tend to be followed by long (short) durations in financial markets.[23] In order to account for this the clustering phenomenon, Engle and Russell (1998) introduced the so called Autoregressive Conditional Duration (ACD) framework to model inter-event durations.

In general, the clustering phenomenon has been ascribed to asymmetric information and interpreted using models developed by Admati and Pfleiderer (1988) or Easley and O'Hara (1992). But Engle and Russell (1998) find that short durations have no effect on price movements when the bid-ask spread is small and conclude (p.1158): "*this suggests that both liquidity-and information-based clustering of transaction rates occur.*" Our model provides a liquidity-based explanation of clustering.

To see why, consider, for instance, an increase in the number of market-takers (more generally, any factor that directly affects the market-takers but not directly the market-makers). The direct effect of this shock is to increase market-takers' aggregate monitoring level and therefore to reduce the average duration from a quote to a trade, $\mathcal{D}_t$. But, in turn, market-makers monitor the market more intensively because makers' and takers' monitoring decisions are complements. Thus, there is also a decline in the duration from a trade to a quote, $\mathcal{D}_m$. This reasoning implies that fluctutations in the number of market-takers (e.g., during the day) create a positive correlation in durations between events ($\mathcal{D}_t$ and $\mathcal{D}_m$) and therefore a clustering in durations between trades.

Thus, clustering in our model is a consequence of the complementarity in monitoring decisions between market-makers and market-takers. For this reason, it would be interesting to directly test whether market-makers' and market-takers' reaction times are complements (the "complementarity hypothesis"), that is, to test whether an increase in $\mathcal{D}_m$ has a positive effect on $\mathcal{D}_t$ and vice versa. Testing this hypothesis is problematic, since inter-event durations are endogenous and simultaneously determined in equilibrium. However, our model suggests several variables that can serve as instruments to identify the causal effect of $\mathcal{D}_m$ on $\mathcal{D}_t$ and vice versa. For

[23] See for instance Hasbrouck (1999) and Pacurar (2006).

instance, an increase in the number of market-takers has a direct negative impact on $\mathcal{D}_t$ but no direct impact on $\mathcal{D}_m$. More generally, the effect of the exogenous variables of the model on, say, $\mathcal{D}_t$, holding $\mathcal{D}_m$ *fixed*, can be obtained from market-takers' best response functions. Using this observation, we obtain the following corollary.

Corollary 7 *There exist two functions $f(\cdot)$ and $g(\cdot)$ such that we can write $\mathcal{D}_m = f(\mathcal{D}_t; \beta, M, c_m)$ and $\mathcal{D}_t = g(\mathcal{D}_m; \gamma, N, c_t)$. Furthermore,*

1. *$\mathcal{D}_m$ and $\mathcal{D}_t$ are complements: $f(\cdot; \beta, M, c_m)$ is increasing in $\mathcal{D}_t$ and $g(\cdot; \gamma, N, c_t)$ is increasing in $\mathcal{D}_m$.*

2. *$f(\mathcal{D}_t; \cdot)$ is increasing in β, and c_m, and decreasing in M, and $g(\mathcal{D}_m; \cdot)$ is increasing in γ, and c_t, and decreasing in N.*

Thus, for a fixed value of $\mathcal{D}_t$, $\mathcal{D}_m$ is determined by c_m, β and M. But, for a fixed value of $\mathcal{D}_m$, these parameters do not affect $\mathcal{D}_t$. In a symmetric way, parameters c_t, γ and N affect $\mathcal{D}_t$ but do not (directly) affect $\mathcal{D}_m$. Thus, empirically, trading fees or the number of participants on either side could be used as instrumental variables to identify the effect of $\mathcal{D}_m$ on $\mathcal{D}_t$ (and $\mathcal{D}_t$ on $\mathcal{D}_m$) using a simultaneous equations approach. For instance, time-variations in the number of market-takers should generate variations in market-takers' average reaction time that are independent of variations in market-makers' average reaction time. Similarly, cross-sectional or time-series variations in make fees should directly affect the response rate of market-makers, but not of market takers.[24]

5.2 Make/Take Fees and the Velocity Ratio

The model has cross-sectional implications for the velocity ratio, $\mathcal{V} \equiv \frac{\mathcal{D}_t}{\mathcal{D}_m} = \frac{\bar{\lambda}}{\bar{\mu}}$. Actually, Corollary 3 implies that, for fixed fees, this ratio should increase in $q = M/N$ and $r = \frac{\gamma}{\beta}$. These implications also hold when fees are set at their optimal level. For instance, using Proposition 4, we obtain that in the thick market case:

$$\mathcal{V}(r, q, c_m^*, c_t^*) = (\frac{\pi_m^* rq}{\pi_t^*})^{1/2} = (rq)^{2/3}. \tag{24}$$

[24] Coopejans et al. (2001) model the dynamic feedback between liquidity suppliers and liquidity demanders using a VAR model for depth on both sides of the limit order book. Our model suggests an alternative approach which consists in quantifying cross-side complementarities using the speed of reactions of each side.

The optimal make-take spread is also positively related to r and q (see Section 4). Thus, if fees are set optimally, the model implies a positive correlation between the make-take spread and the ratio of inter-event durations, $\mathcal{V}$. This prediction is interesting as the make-take spread varies (i) across securities for a given trading platform (see Table 1 in the introduction) and (ii) across trading platforms, for a given security (in which case q may differ across platforms). These variations provide a way to test whether the make-take spread co-varies positively with the velocity ratio, $\mathcal{V}$.

5.3 Tick size, Make-Take Fees, and Trading Rate

Tick size and Make-Take Fees. The model implies a positive association between the make/take spread and the tick size (see Corollary 5 and Figure 4a). Pricing schemes used by trading platforms appear consistent with this implication. Indeed, liquidity rebates for market-makers on the NYSE and Nasdaq, for instance, coincide with the implementation of penny pricing on these markets. Moreover, liquidity rebates were introduced by ECNs such as Archipelago and Island in the 90s which, at that time, were operating on much finer grids than their competitors (Nasdaq and NYSE). Since January 2007, various options markets have implemented pilot programs to quote and trade certain options in pennies ("The Penny Pilot"). For these options, a few trading platforms (e.g., NYSE Arca Options and the Boston Options Exchange) now offer rebates to liquidity providers, as implied by the model when the tick size becomes small.[25] The model makes the additional prediction that the make-take spread should be positively related to the relative size of the market-making sector, q and the relative monitoring cost for the market-taking sector, r (see Corollary 5 and Figures 4b and 4c).

Trading Rate and the Tick Size. The model also implies that, for fixed fees, the trading rate peaks for a strictly positive tick size. Thus, for fixed fees, the relationship between the trading rate and the tick size is non monotonic. In contrast when fees adjust to changes in the tick size, there should be no relationship between the tick size and the trading rate. To see this, let $c_m^*(L, q, r)$ be the optimal make fee for the platform when $\Delta = L$.

[25] See "*Options maker-taker markets gain steam*," Traders Magazine, October 2007.

Corollary 8 *For fixed trading fees, the tick size that maximizes the trading rate is:*

$$\Delta^* = 2(c_m - c_m^*(L, q, r)) + L > 0.$$

Thus, the optimal tick size for the trading platform increases in the fee charged on market-makers (c_m). Moreover it decreases in the number of market-makers relative to the number of market-takers (q) and in market-takers' monitoring cost relative to market-makers' monitoring cost (r). In contrast, if the fees are set optimally, then a change in the tick size has no effect on the trading rate.

Thus, for fixed trading fees, a change in the tick size should affect the trading rate but the direction of the effect is ambiguous. Indeed, other things equal, a higher tick size translates into a higher expected profit for market-makers and a smaller expected profit for market-takers. Thus, an increase in the tick size is conducive to more monitoring by market-makers but less monitoring by market-takers (even though they expect more frequent trading opportunities). When $\Delta < \Delta^*$, the first effect dominates and therefore the trading rate increases in the tick size. In contrast the second effect dominates when $\Delta > \Delta^*$ and therefore the trading rate decreases with the tick size. The model also implies that a decrease in the trading rate after a reduction in the tick size is more likely for stocks for which q is small (since Δ^* is large when q is small). Such a decrease is also more likely when the trading fee for market-makers is relatively high since Δ^* increases in c_m.[26]

Finally, a change in the tick size becomes neutral if the trading platform can freely respond to this change by adjusting its fees (second part of the corollary). The reason is that make/take fees and the tick size are two alternative ways to control how trading gains are split between makers and takers. In the model, these two instruments are perfect substitutes and therefore the maximum rate of trading is independent of whether it is achieved by setting the tick size at Δ^* or by adjusting fees when the tick size deviates from Δ^*. Thus, the model makes the sharp prediction that the effect of a change in the tick size on the trading rate should be very different if trading fees adjust to this change or not.

The tick size can be mandated by regulation or chosen by the trading platform. The tick size (as a percentage of the stock price) for a stock also changes after a stock

[26] Chakravarty et al. (2004) find a significant drop in the trading frequency for all trade sizes categories after the implementation of decimal pricing on the NYSE.

split since such a split leads to a decrease in the nominal stock price.[27] Thus, the previous corollary implies that the trading rate should change after splits, unless the trading platform adjusts fees to neutralize the effect.

5.4 Effects of Algorithmic Trading

In this section, we discuss the effects of algorithmic trading on the trading rate, the bid-ask spread and welfare. Recall that we associate algorithmic trading with smaller monitoring costs (i.e., smaller values of γ and β). All our findings hold even if γ and β are very small. Thus, our model can account for very short delays in traders' reaction to the state of the market as is increasingly observed. For instance, if $M = 20, N = 100$, $c_m = c_t = 0.05$, $\beta = 0.1, \gamma = 0.5$, $\Delta = L = 1$ then in equilibrium $\bar{\lambda} \approx \bar{\mu} \approx 44$. If time is measured in seconds, this means that average inter-events time is about 23 milliseconds.

Trading Volume and Algorithmic Trading. Corollary 1 shows that a decrease in the monitoring cost for market-makers or market-takers triggers an increase in the trading rate. Thus, algorithmic trading should be associated with an increase in the trading rate. This association can be particularly strong because it is magnified by the complementarity in monitoring decisions between market-makers and market-takers. To see this point, consider Figure 5.

This figure shows how a reduction in monitoring costs for market-makers (β) affects the trading rate $\mathcal{R}$, for fixed values of the other parameters ($M = N = 10$, $L = \Delta = 1$, $\gamma = 1$, and $c_m = c_t = 0.05$). The solid curve depicts the equilibrium trading rate when $1/\beta$ increases from 0.5 to 2 (β decreases from 2 to 0.5). It accounts for the fact that a decrease in β leads to more monitoring by market-makers, which prompts market-takers to monitor more and therefore amplifies the initial effect of β on the trading rate through a chain reaction similar to that described in Figure 2. In contrast, the dotted curve shows the evolution of the trading rate when market-takers' aggregate monitoring level is fixed at its equilibrium level when $\beta = 2$, i.e., when we ignore the complementarity between the two sides.

In both cases, the trading rate increases when market-makers' monitoring cost

[27] Angel (1997) develops a theory of the optimal tick size for a firm based on this observation. In his model, firms choose their tick size to minimize their cost of capital. In our model, a change in the tick size may also affect a firm value if this value is positively related to the trading rate. This link is plausible (see Duffie et al. (2005) for instance) but beyond the scope of our model.

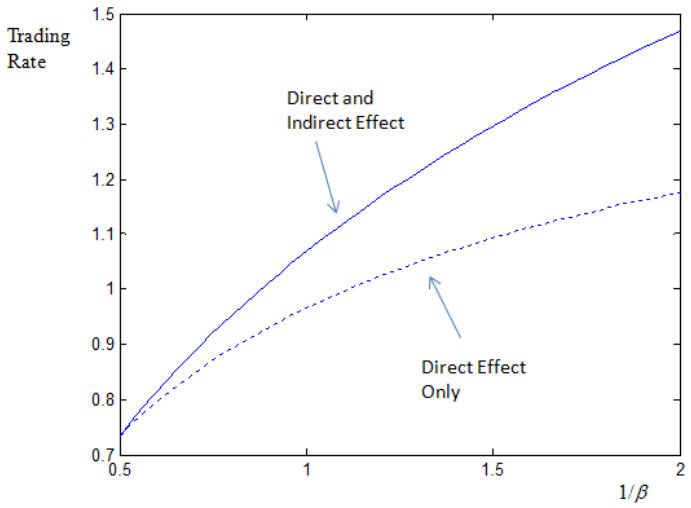

Figure 5: Algorithmic Trading and the Trading Rate

decreases. But the trading rate increases at a much faster rate in equilibrium because the complementarity between market-makers and market-takers amplifies the effect of a decrease in market-makers' monitoring cost. This amplification effect may explain why the rate of increase in trading volume has been so steep in the recent years.[28] For instance, from 2005 to 2007, the number of shares traded on the NYSE rose by 111%, despite the fact that NYSE market share has declined over the same period. This evolution is mainly driven by an increase in the trading rate since the size of trades has steadily declined in the recent years. Our model suggests that the reduction in monitoring costs, combined with cross-side externalities, is one possible cause for this evolution.

Bid-Ask Spread and Algorithmic Trading. A natural question is whether the growth of algorithmic trading will result in tighter bid-ask spreads, as these are often used as measures of market liquidity. The best ask price is either competitive (equal to $a-v_0$) or not competitive (equal to $a+\Delta-v_0$).[29] In each cycle, the bid-ask spread

[28] In this discussion, we have taken the fees as being fixed. It can be checked, using the enveloppe theorem, that the conclusions are unchanged if the fees are set optimally. In fact, the effect of a reduction in β is then even stronger since for each value of β, the trading platform adjusts its fees so as to maximize the trading rate.

[29] Recall that a large number of shares is offered for sale at price $a+\Delta$ by a fringe of competitive traders.

is competitive for an average duration $\mathcal{D}_t$ and uncompetitive for an average duration $\mathcal{D}_m$. Thus, the average half bid-ask spread (denoted ES) is:

$$ES = \theta a + (1-\theta)(a+\Delta) - v_0 = \frac{\Delta}{2} + (1-\theta)\Delta. \tag{25}$$

where

$$\theta \equiv \frac{\mathcal{D}_t}{\mathcal{D}_m + \mathcal{D}_t} = \frac{\mathcal{V}}{1+\mathcal{V}}. \tag{26}$$

The average bid-ask spread decreases with the velocity ratio $\mathcal{V} = \frac{\mathcal{D}_t}{\mathcal{D}_m} = \frac{\bar{\lambda}}{\bar{\mu}}$. Indeed, an increase in this ratio means that, on average, liquidity is supplied at price a more quickly than it is consumed. The result is a smaller bid-ask spread on average.

In equilibrium, the velocity ratio increases in the relative monitoring cost ratio, $r = \frac{\gamma}{\beta}$ (see Corollary 3). Hence, a decrease in β reduces the bid-ask spread whereas a decrease in γ widens the bid-ask spread. Thus, the effect of algorithmic trading on the bid ask spread depends on whether it makes market-makers or market-takers relatively faster. In the former case, algorithmic trading should be associated with a decline in the bid-ask spread whereas in the latter case, it should lead to an increase in the bid-ask spread.

This logic also implies that the impact of faster trading on the bid-ask spread depends on whether trading is faster because market-makers respond more quickly to trades or because market-takers respond more quickly to good prices. Indeed, a decrease in β or γ leads both to a faster market (the trading rate increases). But in the first case, the average quoted bid-ask spread declines whereas in the second case the average quoted bid-ask spread increases. Thus, it is not the speed of trading that affects the bid-ask spread, but the velocity ratio.

Hendershott, Jones, and Menkveld (2009) consider a change in the organization of the NYSE that made algorithmic trading easier for liquidity suppliers. They find empirically that this change reduces the quoted bid-ask spread, as predicted by our model. In contrast, Hendershott and Moulton (2009) study a change in the organization of the NYSE that increases the speed at which liquidity demanders can react to new quotes (i.e., that reduces $\mathcal{D}_t$). They find an increase in the quoted bid-ask spread following this event, which again is in line with the logic of the model.

Welfare and algorithmic trading. Algorithmic trading is often portrayed as socially useless as the gains of algorithmic traders are obtained at the expense of other traders (see for instance Krugman (2009)). To study this question, we analyze

the effect of a decrease in traders' monitoring costs (β or γ) on (i) each participant's expected profits and (ii) on their aggregate expected profit (W). Using equations (8), (9), and (10), this aggregate profit is:

$$\begin{aligned} W(\gamma, \beta, c_m, c_t, M, N) &\equiv \textstyle\sum_{i=1}^{M} \Pi_{im} + \sum_{j=1}^{N} \Pi_{jt} + \Pi_e \\ &= \mathcal{R}\left(\bar{\lambda}^*, \bar{\mu}^*\right) \cdot L - M \cdot C_m(\lambda_1^*) - N \cdot C_t(\mu_1^*). \end{aligned}$$

Thus, *other things being equal*, market participants' aggregate welfare increases in the trading rate. Corollary 1 shows that a decrease in traders' monitoring costs on either side results in a higher trading rate. For this reason, algorithmic trading (a decrease in β or γ) is socially beneficial in our set-up, as shown in the next corollary. The result is indeed even stronger since, other things equal, the expected profit of each participant increases when β or γ decreases. Thus, algorithmic trading results in a Pareto improvement *when trading fees are fixed.*

Corollary 9 *For fixed trading fees, the total expected profit of each participant (market-makers, market-takers, and the trading platform) and therefore aggregate welfare increases when β or γ decreases.*

These results hold for fixed fees. When fees are optimally set by the platform, the welfare analysis is more complex. It is still the case that aggregate welfare increases with algorithmic trading but, counter-intuitively, the side with declining monitoring costs can sometimes be worse off.

To see why, suppose that β decreases. Following this change, the trading platform adjusts its fees, then (i) the make fee increases, (ii) the take fee decreases and (iii) the total fee is unchanged (see Section 4). In choosing optimally its fees, the trading platform amplifies the effect of the decrease in β on the trading rate. Thus, it gets a larger expected profit. As the total fee is unchanged and the take fee is smaller, market-takers are also clearly better off. But this gain is obtained at the expense of the market-makers since they end up paying a larger fee. The net effect of a decrease in β on market-makers' welfare is therefore ambiguous when fees are endogenous.

This point can be easily seen when $M = N = 1$. In this case, in equilibrium, the

market-maker's expected profit per unit of time when fees are set optimally is:[30]

$$\Pi_{im}(\beta,\gamma) = \frac{(L-\overline{c})^2\beta^{\frac{1}{4}}(\beta^{\frac{1}{4}}+2\gamma^{\frac{1}{4}})}{4(\beta^{\frac{1}{4}}+\gamma^{\frac{1}{4}})^6}.$$

Thus, the market-maker's expected profit decreases in γ. That is, a reduction in the market-taker's monitoring cost always makes the market-maker better off. In contrast, the market-maker's expected profit is non monotonic in β and goes to zero when β goes to zero. Thus, a decline in the market-maker's monitoring cost can make her worse off. Actually, in this case, the trading platform charges a higher fee on the market-maker. As a consequence, part of the cost reduction in monitoring is transferred to the market-taker. Symmetric results are obtained for market-takers.

It is worth stressing two limitations here. First, as explained previously, algorithmic trading can sometimes result in a higher average bid-ask spread. This effect does not affect market-takers' welfare in our set up because they only trade when the bid-ask spread is small. In reality, some market-takers might be willing to trade at worse prices than the competitive quotes and these traders will be hurt when the bid-ask spread becomes larger on average.[31] Second, we do not model why investors want to trade (L is exogenous). For these reasons, our welfare results must be interpreted cautiously. Yet, they suggest that accounting for make/take fees is important for welfare analyses of changes in market structure.

6 Conclusion

In this paper, we develop a model of trading in which traders react with a delay to profit opportunities because monitoring the market is costly. One group of traders ("market-makers") specializes in posting quotes while another group of traders ("market-takers") specializes in hitting quotes. Each market-maker monitors the market to be the first to submit the competitive quote after a transaction. Each market-taker monitors the market to be the first to hit the competitive quote. In this

[30] Detailed calculations are skipped for brevity. They can be obtained upon request. We provide the expression for market-makers' expected profit for an arbitrary value of the total fee charged by the platform. This does not affect our conclusion here since the optimal total fee in this case is $\overline{c} = L/2$, as in the thick market case, and is therefore independent of β and γ .

[31] One way to introduce this effect in the model is to assume that there are market-takers with a fixed (exogenous) monitoring intensity who are willing to buy the security at a or $a+\Delta$. These traders are hurt if the best offer is more likely to be $a+\Delta$ when they inspect the market.

way, we model the high frequency make/take liquidity cycles observed in electronic limit order markets.

In our theory, the monitoring decisions of market-makers on the one hand and market-takers on the other hand are self-reinforcing. This feature has several implications. First, it creates a coordination problem between market-makers and market-takers that results in multiple equilibria with differing levels of trading activity. Second, it implies that the speed at which market-takers consume liquidity is positively related to the speed at which market-makers supply liquidity and vice versa. This property offers a new explanation for the clustering in durations between trades.

The theory also has implications for make/take fees. These fees affect the trading rate because they determine how gains from trade are split between market-makers and market-takers and thereby the incentive of each side to respond promptly to a change in the state of the market. As a consequence, there is a breakdown of the total fee between the two sides that maximizes the trading rate. In particular, it is optimal for the trading platform to charge lower fees on market-makers when (i) the tick size is low, (ii) the number of market-makers relative to the number of market-takers is low or (iii) the monitoring cost of market-takers relative to the monitoring costs of market-makers is low.

We also use the model to study the effect of algorithmic trading (that we interpret as a reduction of monitoring costs). The model implies that algorithmic trading should lead to a sharp increase in the trading rate. Moreover, it should lead to a decrease in the bid-ask spread if and only if it increases the speed of reaction of market-makers relative to the speed of reaction of market-takers (the "velocity ratio"). Last, algorithmic trading is socially beneficial because it increases the rate at which gains from trade are realized. Yet, adjustments in trading fees redistribute the social gain of algorithmic trading between participants. For this reason, automation of one side may, counter-intuitively, make this side worse off after adjustments in make/take fees.

The model could be extended in many directions. First, in reality, market-makers are exposed to adverse changes in the value of the security. They can reduce this risk by monitoring the flow of information and cancel their quotes when new information arrives (see Foucault et al. (2003) for a theoretical analysis). It would be interesting to incorporate this possibility in the model. Intuitively, exposure to informed trading reduces market-makers' expected profits and should therefore lead to higher make

fees, other things equal. Second, in our model, traders do not choose the side on which they are active. An interesting extension would be to endogenize the number of makers and takers by allowing traders to choose their side. Last, our model considers a single trading platform. In reality, securities often trade on multiple platforms. The economic forces analyzed in our paper should still hold in a multi-market environment as long as monitoring is costly. In particular, the make/take fees charged by a platform should still affect its trading rate as they will affect makers and takers' incentives to monitor the platform. Yet, inter-market competition may add other considerations to the choice of make/take fees. Moreover, considering the multi-market environment may allow to model the opportunity cost of monitoring, which is exogenous in our set-up.

7 Appendix

Proof of Proposition 1: Direct from the argument in the text.

Proof of Proposition 2: From (4) and (8), the first order condition for market-maker i is:

$$\frac{\bar{\mu}\left(\bar{\mu}+\bar{\lambda}-\lambda_i\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_m}{\beta}=\lambda_i. \tag{27}$$

Summing over all $i=1,\ldots M$, we obtain

$$\frac{\bar{\mu}\left(\left(\bar{\mu}+\bar{\lambda}\right)M-\bar{\lambda}\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_m}{\beta}=\bar{\lambda}. \tag{28}$$

Similarly, for market-takers we obtain,

$$\frac{\bar{\lambda}\left(\left(\bar{\mu}+\bar{\lambda}\right)N-\bar{\mu}\right)}{\left(\bar{\lambda}+\bar{\mu}\right)^2}\frac{\pi_t}{\gamma}=\bar{\mu}. \tag{29}$$

Let $\Omega\equiv\frac{\bar{\lambda}}{\bar{\mu}}$. Dividing the left-hand-side of (28) and (29) by $\bar{\mu}^2$ we obtain,

$$\frac{M+(M-1)\,\Omega}{(1+\Omega)^2}\frac{\pi_m}{\beta} = \bar{\lambda}. \tag{30}$$

$$\frac{\Omega\left((1+\Omega)\,N-1\right)}{(1+\Omega)^2}\frac{\pi_t}{\gamma} = \bar{\mu} \tag{31}$$

Dividing these two equations gives,

$$\frac{(M+(M-1)\,\Omega)}{\Omega^2\left((1+\Omega)\,N-1\right)}z=1, \tag{32}$$

or equivalently,

$$\Omega^3 N + (N-1)\Omega^2 - (M-1)\, z\Omega - Mz = 0.$$

We argue that this cubic equation has a unique positive solution. Indeed, this equation is equivalent to

$$\Omega = g(\Omega, M, N, z). \tag{33}$$

with

$$g(\Omega, M, N, z) = \frac{(M-1)z}{\Omega N} + \frac{Mz}{N\Omega^2} - \frac{N-1}{N}. \tag{34}$$

Function $g(\cdot, M, N, z)$ decreases in Ω. It tends to plus infinity as Ω goes to zero, and to $-\frac{N-1}{N}$ as Ω goes to infinity. Thus, (33) has a unique positive solution that we denote by Ω^*. We obtain the aggregate monitoring levels in equilibrium by inserting this root into Equations (30) and (31).

Now note that the equilibrium trading strategies must symmetric among the market-makers and market-takers. That is, $\lambda_1^* = \lambda_2^* = \ldots = \lambda_M^*$ and $\mu_1^* = \mu_2^* = \ldots = \mu_N^*$.[32] Hence, in equilibrium $\lambda_i = \bar{\lambda}/M$ and $\mu_j = \bar{\mu}/N$ for all i, j. This completes the proof. ■

Proof of Corollary 1: Recall that Ω^* is such that:

$$\Omega^* = g(\Omega^*, M, N, z), \tag{35}$$

where $g(\cdot)$ is defined in equation (34). It is immediate that $g(\cdot)$ increases in M, decreases in N, and increases in z. As $g(\cdot)$ decreases in Ω, we have

$$\frac{\partial \Omega^*}{\partial M} > 0, \tag{36}$$

$$\frac{\partial \Omega^*}{\partial N} < 0. \tag{37}$$

$$\frac{\partial \Omega^*}{\partial z} > 0 \tag{38}$$

Now, using Equations (36) and (12), we conclude that:

$$\frac{\partial \lambda_i^*}{\partial N} = \frac{-\frac{\partial \Omega^*}{\partial N} \cdot ((M+1) + (M-1)\Omega^*)}{(1+\Omega^*)^3} \left(\frac{\pi_m}{M\beta}\right) > 0.$$

[32] Indeed, suppose for example that $\lambda_1^* > \lambda_2^*$. Then, from (27),

$$\frac{\bar{\mu}\left(\bar{\mu} + \bar{\lambda} - \lambda_1^*\right)}{\left(\bar{\lambda} + \bar{\mu}\right)^2} \frac{\pi_m}{\beta} > \frac{\bar{\mu}\left(\bar{\mu} + \bar{\lambda} - \lambda_2^*\right)}{\left(\bar{\lambda} + \bar{\mu}\right)^2} \frac{\pi_m}{\beta},$$

which simplifies to $\lambda_1^* < \lambda_2^*$ - a contradiction.

Hence, $\frac{\partial \bar{\lambda}^*}{\partial M} > 0$. Similarly, using equations (37) and (13), we deduce that

$$\frac{\partial \mu_j^*}{\partial M} > 0. \tag{39}$$

Hence, $\frac{\partial \bar{\mu}^*}{\partial M} > 0$. We also have

$$\Omega^* = \frac{\bar{\lambda}^*}{\bar{\mu}^*}.$$

Thus, using equations (36) and (37), we conclude that $\frac{\bar{\lambda}^*}{\bar{\mu}^*}$ increases in M and decreases in N. Equation (39) implies that $\bar{\mu}^*$ increases in M. Thus it must be the case that $\bar{\lambda}^*$ increases in M as well. A similar argument shows that $\bar{\mu}^*$ increases in N.

Now, consider the effect of a change in β on market-takers' monitoring intensities. We have (see Proposition 2),

$$\mu_j^* = \zeta(\Omega^*)\left(\frac{\pi_t}{N\gamma}\right),$$

where

$$\zeta(\Omega^*) = \left(\frac{\Omega^*\left(\left(1+\Omega^*\right)N-1\right)}{\left(1+\Omega^*\right)^2}\right).$$

Thus

$$\frac{\partial \mu_j^*}{\partial \beta} = \left(\frac{\partial \zeta(\Omega^*)}{\partial \Omega^*}\frac{\partial \Omega^*}{\partial z}\frac{\partial z}{\partial \beta}\right)\left(\frac{\pi_t}{N\gamma}\right)$$

We have $\frac{\partial \zeta(\Omega^*)}{\partial \Omega^*} > 0$. Moreover $\frac{\partial \Omega^*}{\partial z} > 0$ and $\frac{\partial z}{\partial \beta} < 0$. Thus

$$\frac{\partial \mu_j^*}{\partial \beta} < 0,$$

which implies that $\frac{\partial \bar{\mu}^*}{\partial \beta} < 0$. Now, since $\bar{\lambda}^* = \Omega^* \bar{\mu}^*$, we have:

$$\frac{\partial \bar{\lambda}^*}{\partial \beta} = \Omega^* \frac{\partial \bar{\mu}^*}{\partial \beta} + \frac{\partial \Omega^*}{\partial z}\frac{\partial z}{\partial \beta}\bar{\mu}^* < 0,$$

which implies $\frac{\partial \lambda_j^*}{\partial \beta} < 0$. The impact of make/take fees on traders' aggregate monitoring levels of is obtained in the same way. The second part of the corollary directly follows from the first part and the definition of the trading rate (equation (4)).

Proof of Corollary 2: Recall that $\frac{\bar{\lambda}^*}{\bar{\mu}^*} = \Omega^*$. Using equation (14), it is readily checked that $\Omega^* = 1$ if and only if $z = \frac{2N-1}{2M-1}$. Thus, $\bar{\lambda}^* = \bar{\mu}^*$ if and only if $z = \frac{2N-1}{2M-1}$. Moreover, as shown in the proof of Corollary 1, Ω^* increases in z. Hence, $\bar{\lambda}^* > \bar{\mu}^*$ iff $z > \frac{2N-1}{2M-1}$. ■

Proof of Corollary 3: Recall that $\mathcal{V} \equiv \frac{\mathcal{D}_t}{\mathcal{D}_m} = \frac{\bar{\lambda}^*}{\bar{\mu}^*} = \Omega^*$. We know from the proof of Corollary 1 that Ω^* increases in z and M and decreases in N. The corollary is then immediate using the definition of z (equation (11)). ■

Proof of Corollary 4: Using Proposition 2, we deduce that:

$$\begin{aligned}\lim_{M\to\infty} \lambda_i^* &= \lim_{M\to\infty} \left(\frac{M+(M-1)\,\Omega^*}{M\left(1+\Omega^*\right)^2}\right)\left(\frac{\pi_m}{\beta}\right) = \lim_{M\to\infty} \left(\frac{1+\frac{M-1}{M}\Omega^*}{\left(1+\Omega^*\right)^2}\right)\left(\frac{\pi_m}{\beta}\right) \\ &= \frac{1}{1+\Omega^\infty}\left(\frac{\pi_m}{\beta}\right) = \frac{1}{1+(zq)^{\frac{1}{2}}}\frac{\pi_m}{\beta} \quad \text{(using (15))}.\end{aligned}$$

A similar argument is used to derive μ_j^∞. ■

Proof of Proposition 3: We have

$$\begin{aligned}\frac{\partial \mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial c_m} &= \mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)^2\left(\frac{\partial \bar{\lambda}^*}{\partial c_m}\frac{1}{\bar{\lambda}^{*2}} + \frac{\partial \bar{\mu}^*}{\partial c_m}\frac{1}{\bar{\mu}^{*2}}\right) \\ &= \frac{\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)^2}{c_m}\left(\frac{\eta_{mm}}{\bar{\lambda}^*} + \frac{\eta_{tm}}{\bar{\mu}^*}\right). \qquad (40)\end{aligned}$$

and

$$\frac{\partial \mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial c_t} = \frac{\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)^2}{c_t}\left(\frac{\eta_{mt}}{\bar{\lambda}^*} + \frac{\eta_{tt}}{\bar{\mu}^*}\right). \qquad (41)$$

The optimal fee structure is such that

$$\frac{\partial \mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial c_m} = \frac{\partial \mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*)}{\partial c_t}.$$

Thus, using equations (40) and (41), we deduce that

$$\frac{(\bar{\lambda}^*)^{-1}\eta_{mm} + (\bar{\mu}^*)^{-1}\eta_{tm}}{(\bar{\lambda}^*)^{-1}\eta_{mt} + (\bar{\mu}^*)^{-1}\eta_{tt}} = \frac{c_m}{c_t}.$$

Now, (20) follows directly from this equation and the fact that $c_m + c_t = \bar{c}$. ■

Proof of Proposition 4: We fix $q > 0$, and let $N = \frac{M}{q}$. Note that there is a one-to-one mapping between the fees charged by the trading platform and the per trade trading profits obtained by the market-making side and the market-taking side, π_m and π_t. Thus, instead of using c_m and c_t as the decision variables of the platform, we can use π_m and π_t. It turns out that this is easier. Thus, for a fixed $\bar{c}$, we rewrite the platform's problem as:

$$\begin{aligned}&Max_{\pi_m,\pi_t}\mathcal{R}(\bar{\lambda}^*,\bar{\mu}^*) \\ &s.t \;\; \pi_t + \pi_m = L - \bar{c}.\end{aligned}$$

Moreover, using that $\pi_m = L - \bar{c} - \pi_t$, we can present $\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)$ as a function of π_t only. We know that

$$\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*) = \frac{\bar{\lambda}^* \bar{\mu}^*}{\bar{\lambda}^* + \bar{\mu}^*} = \frac{\bar{\lambda}^*}{1 + \Omega^*}. \tag{42}$$

The first order condition with respect to π_t gives

$$\frac{\partial \bar{\lambda}^*}{\partial \pi_t}(1 + \Omega^*) - \frac{\partial \Omega^*}{\partial \pi_t}\bar{\lambda}^* = 0,$$

or equivalently,

$$\frac{\partial \bar{\lambda}^*}{\partial \pi_t} = \mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)\frac{\partial \Omega^*}{\partial \pi_t}.$$

Since $\bar{\lambda}^* = M\lambda_1^*$, we can divided both sides by M and obtain

$$\frac{\partial \lambda_1^*}{\partial \pi_t} = \frac{\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{M}\frac{\partial \Omega^*}{\partial \pi_t}. \tag{43}$$

Since the first order condition holds for any M, we can take limits on both sides to obtain a necessary condition for the large market:

$$\lim_{M\to\infty} \frac{\partial \lambda_1^*}{\partial \pi_t} = \lim_{M\to\infty} \frac{\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{M} \cdot \lim_{M\to\infty} \frac{\partial \Omega^*}{\partial \pi_t}. \tag{44}$$

Straightforward calculations show that

$$\begin{aligned}
\lim_{M\to\infty} \frac{\partial \Omega^*}{\partial \pi_t} &= \frac{d\Omega^\infty}{d\pi_t} = -\frac{q}{2\Omega^\infty}\frac{L - c}{\pi_t^2}\frac{\gamma}{\beta}, \text{ and} \\
\lim_{M\to\infty} \frac{d\lambda_1^*}{d\pi_t} &= -\frac{1}{\beta\left(1 + \Omega^\infty\right)} - \frac{1}{\left(1 + \Omega^\infty\right)^2} \cdot \frac{d\Omega^\infty}{d\pi_t} \cdot \frac{L - \bar{c} - \pi_t}{\beta},
\end{aligned}$$

where Ω^∞ is given by (15). Furthermore, it is direct from (42) that

$$\lim_{M\to\infty} \frac{\mathcal{R}(\bar{\lambda}^*, \bar{\mu}^*)}{M} = \frac{\lambda_1^\infty}{1 + \Omega^\infty},$$

where λ_1^∞ is given by (16). Plugging back into (44) and simplifying yields

$$1 - \frac{q\left(L - \bar{c} - \pi_t\right)}{\Omega^\infty\left(1 + \Omega^\infty\right)} \cdot \frac{L - \bar{c}}{\pi_t^2} \cdot \frac{\gamma}{\beta} = 0,$$

or,

$$1 - \frac{zq}{\left(1 + \Omega^\infty\right)\Omega^\infty}\frac{L - \bar{c}}{\pi_t} = 0, \tag{45}$$

which simplifies to

$$\frac{\pi_t}{L - \bar{c}} = \frac{\Omega^\infty}{1 + \Omega^\infty}. \tag{46}$$

Denote

$$w \equiv \frac{\pi_t}{L - \bar{c}}. \tag{47}$$

Then (46) imposes

$$w = \frac{\Omega^{\infty}}{1 + \Omega^{\infty}} = \frac{1}{1 + (zq)^{\frac{1}{2}}}. \tag{48}$$

Now, observe that

$$z = r\frac{1 - w}{w}.$$

Thus, we can rewrite equation (48) as

$$w = \frac{1}{1 + \left(\frac{1-w}{w}\right)^{-0.5} (rq)^{-0.5}}.$$

It is immediate that this equation has a unique solution:

$$w^* = \frac{(rq)^{\frac{1}{3}}}{1 + (rq)^{\frac{1}{3}}}.$$

From (47) we obtain,

$$\pi_t = \frac{L - \bar{c}}{1 + (rq)^{-\frac{1}{3}}}, \tag{49}$$

and,

$$\pi_m = L - \bar{c} - \pi_t = \frac{L - \bar{c}}{1 + (rq)^{\frac{1}{3}}}. \tag{50}$$

Given this,

$$z = \frac{\frac{L-\bar{c}}{1+(qr)^{\frac{1}{3}}}}{\frac{L-\bar{c}}{1+(qr)^{-\frac{1}{3}}}} r = q^{-\frac{1}{3}} r^{\frac{2}{3}}.$$

And,

$$\Omega^{\infty} = (zq)^{\frac{1}{2}} = (rq)^{\frac{1}{3}}. \tag{51}$$

The optimal fees in the large market are:

$$\begin{aligned} c_m &= \frac{\Delta}{2} - \pi_m = \frac{\Delta}{2} - \frac{L - \bar{c}}{1 + (qr)^{\frac{1}{3}}} \\ c_t &= L - \frac{\Delta}{2} - \pi_t = L - \frac{\Delta}{2} - \frac{L - \bar{c}}{1 + (qr)^{-\frac{1}{3}}}. \end{aligned}$$

Finally, the monitoring intensities in the large market are obtained by plugging these expressions into Corollary 4. ■

Proof of Corollary 5: The result follows directly from equation (21) ■

Proof of Corollary 6: We fix $q > 0$, and let $N = \frac{M}{q}$.For any given M, maximizing $\mathcal{R}(\lambda^*, \mu^*)\bar{c}$ is equivalent to maximizing $\frac{\mathcal{R}(\lambda^*,\mu^*)}{M}\bar{c}$, which in turn (using (42) and that $\bar{\lambda}^* = M\lambda_1^*$) is equivalent to maximizing $\frac{\lambda_1^*}{1+\Omega^*}\bar{c}$. Denote $\mathcal{H}(\bar{c}) \equiv \frac{\lambda_1^*}{1+\Omega^*}$. Then, to find the optimal total fee $\bar{c}$ in the large market case we need to find the limit as M tends to infinity of

$$\arg\max_{\bar{c}\geq 0} \mathcal{H}(\bar{c})\,\bar{c}.$$

The FOC for a given M is

$$\mathcal{H}(\bar{c}) + \mathcal{H}'(\bar{c})\,\bar{c} = 0. \tag{52}$$

Note that $\mathcal{H}$ depends on $\bar{c}$ only through its dependence on λ_1^* and Ω^*. It follows that

$$\mathcal{H}'(\bar{c}) = \frac{\partial \mathcal{H}}{\partial \lambda_1^*}\frac{\partial \lambda_1^*}{\partial \bar{c}} + \frac{\partial \mathcal{H}}{\partial \Omega^*}\frac{\partial \Omega^*}{\partial \bar{c}} = \frac{1}{1+\Omega^*}\frac{\partial \lambda_1^*}{\partial \bar{c}} - \frac{\lambda_1^*}{(1+\Omega^*)^2}\frac{\partial \Omega^*}{\partial \bar{c}}. \tag{53}$$

Since (52) holds for any M, we can take the limit as $M \to \infty$. We have,

$$\lim_{M\to\infty} \mathcal{H}(\bar{c}) = \frac{\lambda_1^\infty}{1+\Omega^\infty} = \frac{L-\bar{c}}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^3} \quad \text{(using (23) and (51))}.$$

It can also be verified using (23) and (51) that

$$\begin{aligned}
\lim_{M\to\infty} \frac{\partial \lambda_1^*}{\partial \bar{c}} &= \frac{\partial \lambda_1^\infty}{\partial \bar{c}} = -\frac{1}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^2}, \text{ and} \\
\lim_{M\to\infty} \frac{\partial \Omega^*}{\partial \bar{c}} &= \lim_{M\to\infty} \frac{\partial \Omega^\infty}{\partial \bar{c}} = 0.
\end{aligned}$$

Thus, from (53),

$$\lim_{M\to\infty} \mathcal{H}'(\bar{c}) = -\frac{1}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^3}.$$

And, in the limit (52) becomes

$$\frac{L-\bar{c}}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^3} - \frac{1}{\beta\left(1+(qr)^{\frac{1}{3}}\right)^3}\bar{c} = 0,$$

which gives $\bar{c} = \frac{L}{2}$. ■

Proof of Corollary 7: Using equation (30) in the proof of Proposition 2, we deduce that for a fixed $\bar{\mu}$, market-makers' aggregate monitoring level, $\bar{\lambda}$ solve

$$F(\bar{\lambda}; \bar{\mu}, \beta, c_m, M) = 0, \tag{54}$$

where,

$$F(\bar{\lambda};\bar{\mu},\beta,c_m,M) \equiv M + \frac{(M-1)\pi_m\bar{\lambda}}{\bar{\mu}} - \beta\bar{\lambda}(1+\frac{\bar{\lambda}}{\bar{\mu}})^2.$$

It is easily shown that for all parameter values, equation (54) has a unique positive solution $\bar{\lambda}$. Let $\bar{\lambda} = \varphi(\bar{\mu};\beta,c_m,M)$ be this solution. Now, using the implicit function theorem:

$$\frac{d\varphi}{d\bar{\mu}} = -\frac{\frac{\partial F}{\partial \bar{\mu}} \mid_{\bar{\lambda}=\varphi(\bar{\mu};\beta,c_m,M)}}{\frac{\partial F}{\partial \bar{\lambda}} \mid_{\bar{\lambda}=\varphi(\bar{\mu};\beta,c_m,M)}}.$$

Using the expression for $F(\cdot)$, we obtain $\frac{\partial F}{\partial \bar{\mu}} \mid_{\bar{\lambda}=\varphi(\bar{\mu};\beta,c_m,M)} > 0$ and $\frac{\partial F}{\partial \bar{\lambda}} \mid_{\bar{\lambda}=\varphi(\bar{\mu};\beta,c_m,M)} < 0$. Thus:

$$\frac{d\varphi}{d\bar{\mu}} > 0 \tag{55}$$

Now, since we have $\bar{\lambda} = \varphi(\bar{\mu};\beta,c_m,M)$, we deduce that:

$$\mathcal{D}_m = f(\mathcal{D}_t;\beta,M,c_m),$$

with $f(\mathcal{D}_t;\beta,M,c_m) = \frac{1}{\varphi(\frac{1}{\mathcal{D}_t};\beta,c_m,M)}$. Then, from equation (55), we deduce that $\frac{\partial f}{\partial \mathcal{D}_t} > 0$. The result for g is established in a parallel manner. This completes the first part of the corollary. The second part is obtained using a similar arguments (again applying the implicit function theorem). We omit the details for brevity. ■

Proof of Corollary 8: Define $\hat{c}_m = \frac{L}{2} - \frac{\Delta}{2} + c_m$ and $\hat{c}_t = \overline{c} - \hat{c}_m = \frac{\Delta}{2} - \frac{L}{2} + c_t$. Observe that we can write market-makers' and market-takers' payoffs as:

$$\begin{aligned}
\Pi_{im}(\lambda_i;\hat{c}_m) &= \frac{\lambda_i\bar{\mu}\left(\frac{\Delta}{2}-c_m\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\lambda_i^2 = \frac{\lambda_i\bar{\mu}\left(\frac{L}{2}-\hat{c}_m\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\lambda_i^2 \\
\Pi_{jt}(\mu_j;\hat{c}_t) &= \frac{\mu_j\bar{\lambda}\left(L-\frac{\Delta}{2}-c_t\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\mu_j^2 = \frac{\mu_j\bar{\lambda}\left(\frac{L}{2}-\hat{c}_t\right)}{\bar{\lambda}+\bar{\mu}} - \frac{1}{2}\beta\mu_j^2
\end{aligned}$$

These payoffs are those obtained when $\Delta = L$ and fees are set at $\hat{c}_m$ and $\hat{c}_t$. Thus, the values of $\hat{c}_m$ and $\hat{c}_t$ that maximize the trading rate are:

$$\begin{aligned}
\hat{c}_m^* &= c_m^*(L,q,r) \\
\hat{c}_t^* &= c_t^*(L,q,r).
\end{aligned}$$

These values are independent of the tick size. Thus, for arbitrary values of the fees, the trading platform can make sure that the trading rate is maximal by setting a tick size equal to Δ^* such that:

$$\hat{c}_m = \frac{L}{2} - \frac{\Delta^*}{2} + c_m = c_m^*(L,q,r),$$

that is $\Delta^* = 2(c_m - c_m^*(L,q,r)) + L$.

Moreover, *when the fees are set at their optimal values*, traders' expected profits do not depend on the tick size since traders' expected payoffs at the optimal fees do not depend on the tick size (e.g., $\Pi_{im}(\lambda_i; \hat{c}_m^*)$ does not depend on Δ). Thus, their optimal monitoring levels and therefore the trading rate does not depend on the tick size, which proves the second part of the corollary. ■

Proof of Corollary 9: Consider first the aggregate expected profit for market-takers. We have:

$$\Pi_t(\mu_1^*, .., \mu_j^*, ..., \mu_N^*, \overline{\lambda}^*; \gamma, \beta, c_m, c_t) = \textstyle\sum_j \Pi_{jt}(\mu_j^*, \bar{\lambda}^*; \gamma, \beta, M, N)$$

Thus,

$$\begin{aligned}
\frac{d\Pi_t}{d\gamma} &= \textstyle\sum_j \left(\frac{\partial \Pi_{jt}}{\partial \mu_j^*} \frac{\partial \mu_j^*}{\partial \gamma} + \frac{\partial \Pi_{jt}}{\partial \bar{\lambda}^*} \frac{\partial \bar{\lambda}^*}{\partial \gamma} + \frac{\partial \Pi_{jt}}{\partial \gamma} \right) \\
\frac{d\Pi_t}{d\beta} &= \textstyle\sum_j \left(\frac{\partial \Pi_{jt}}{\partial \mu_j^*} \frac{\partial \mu_j^*}{\partial \beta} + \frac{\partial \Pi_{jt}}{\partial \bar{\lambda}^*} \frac{\partial \bar{\lambda}^*}{\partial \beta} + \frac{\partial \Pi_{jt}}{\partial \beta} \right)
\end{aligned}$$

Now, the envelope theorem implies that $\frac{\partial \Pi_{jt}}{\partial \mu_j^*} = 0$ for all j. Moreover, the cross-side complementarity implies $\frac{\partial \Pi_{jt}}{\partial \bar{\lambda}^*} > 0$ for all j, and Corollary 1 yields $\frac{\partial \bar{\lambda}^*}{\partial \gamma} < 0$ and $\frac{\partial \bar{\lambda}^*}{\partial \beta} < 0$. Last, for all j, $\frac{\partial \Pi_{jt}}{\partial \gamma} = -\frac{1}{2}\left(\mu_j^*\right)^2 < 0$ and $\frac{\partial \Pi_{jt}}{\partial \beta} = 0$. Thus, $\frac{d\Pi_t}{d\gamma} < 0$ and $\frac{d\Pi_t}{d\beta} < 0$. This establishes the first part of the proposition for the market-taking side. The proof for the market-makers is parallel. Last, we have proved in Corollary 1 that the trading rate decreases when β or γ increases. It follows that the expected profit of the platform decreases with β or γ. ■

References

[1] Admati, A.R., and Pfleiderer, (1988), "A Theory of Intraday Patterns : Volume and Price Variability," *The Review of Financial Studies*, 1, 3-40.

[2] Angel, J. (1997), "Tick size, share prices and stock splits," *Journal of Finance* 52, 655-680.

[3] Biais, B., Hillion, P., and Spatt, C. (1995), "An empirical analysis of the limit order book and the order flow in the Paris bourse," *Journal of Finance* 50, 1655-1689.

[4] Biais, B. and Weill, P.O. (2008), "Algorithmic Trading and the Dynamics of the Order Book," Manuscript, Toulouse University, IDEI.

[5] Chakravarty, S., Wood, R. and R.A. Van Ness (2004), "Decimals and liquidity: a study of the NYSE," *Journal of Financial Research* 27, 75-94.

[6] Coopejans, M, Domowitz, I. and Madhavan A. (2001), "Liquidity in an automated auction," Working paper, ITG.

[7] Corwin, S. and Coughenour, J.(2008), "Limited attention and the allocation of effort in securities trading," forthcoming in *Journal of Finance*.

[8] Degryse, H., De Jong F., Van Rvenswaaij, M. and Wuyts, G.(2005), "Aggressive orders and the resiliency of a limit order market," *Review of Finance*, 9, 201-242.

[9] Dow, J., (2004), "Is liquidity self-fulfilling?" *Journal of Business* 77, 895-908.

[10] Duffie, D, Gârleanu, N. and Pedersen, L.H (2005), "Over-the-counter markets," *Econometrica* 73, 1815-1847.

[11] Easley, D. and O'Hara, M. (1992), "Time and the Process of Security Price Adjustment," *Journal of Finance* 47, 577-605.

[12] Easley, D., O'Hara, M. and G. Saar (2001), "How Stock Splits Affect Trading: A Microstructure Approach," *Journal of Financial and Quantitative Analysis* 36, 25-51.

[13] Engle, R.F. and J.R. Russell (1998), "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica* 66, 1127-1162.

[14] Foucault, T. and Menkveld, A. (2008), "Competition for order flow and smart order routing systems", *Journal of Finance* 63, 119-158.

[15] Foucault, T., Roëll, A., and Sandås, P. (2003), "Market Making With Costly Monitoring: An Analysis of SOES Trading," *Review of Financial Studies* 16, 345-384.

[16] Foucault, T., Kadan, O., and Kandel, E. (2005), "Limit order book as a market for liquidity," *Review of Financial Studies*, 18, 1171-1217.

[17] Glosten, L. R. (1994), "Is the electronic open limit order book inevitable?" *Journal of Finance* 49, 1127-1161.

[18] Hasbrouck J. (1999), "Trading fast and slow: security markets in real time," mimeo, NYU.

[19] Hasbrouck J. and Saar G. (2009), "Technology and liquidity provision: the blurring of traditional definitions," *Journal of Financial Markets 12,* 143-172.

[20] Hall, A. and Hautsch, N. (2007), "Modelling the buy and sell intensity in a limit order book market," *Journal of Financial Markets* 10, 249-286.

[21] Harris, L. (1990), "Liquidity, trading rules and electronic trading systems," New-York University, Salomon Brothers monograph series.

[22] Hendershott, T., Jones, C. and Menkveld, A. (2009), "Does algorithmic trading improve liquidity," Working Paper, University of California, Berkeley.

[23] Hendershott, T., and P. C. Moulton, (2009), "Speed and stock market quality: The NYSE's hybrid," Working Paper, University of California, Berkeley.

[24] Hollifield, B., Miller, R. A., Sandas, P. (2004), "Empirical analysis of limit order markets," *Review of Economic Studies* 71, 1027-1063.

[25] Kandel, E. and Marx, L. (1999), "Payments for order flow on Nasdaq," *Journal of Finance* 49, 35-66.

[26] Krugman, P. (2009), "Rewarding bad actors," New-York Times, August 2, 2009.

[27] Large, J. (2007), "Measuring the resiliency of an electronic limit order book," *Journal of Financial Markets*, 1-25.

[28] Large, J. (2009), "A market clearing role for inefficiency on a limit order book," *Journal of Financial Economics*, 102-117.

[29] Liu, W. (2007), "Monitoring and limit order submission risks," forthcoming *Journal of Financial Markets*.

[30] Obizhaeva A. and J. Wang (2006), "Optimal Trading Policy and Demand/Supply Dynamics," working paper, MIT.

[31] Pacurar, M. (2006), "Autoregressive Conditional Duration (ACD) models in finance: a survey of the theoretical and empirical literature," working paper, Dalhousie University.

[32] Pagano, M. (1989), "Trading volume and asset liquidity," *Quarterly Journal of Economics*, 104, 255-276.

[33] Parlour, C. (1998), "Price dynamics in limit order markets," *Review of Financial Studies*, 11, 789-816.

[34] Parlour, C and Rajan, U. (2003), "Payment for order flow," *Journal of Financial Economics*, 68, 379-411.

[35] Rochet, JC and Tirole, J.(2006), "Two sided markets: a progress report," *Rand Journal of Economics, 3*7, 645-667.

[36] Ross, S. M., (1996), Stochastic Processes, John Wiley & Sons, Inc.

[37] Roşu, I. (2008), "A dynamic model of the limit order book," *Review of Financial Studies*, forthcoming.

[38] Sandås, P., (2001), "Adverse selection and competitive market making: Empirical evidence from a limit order market," *Review of Financial Studies* 14, 705-734.

[39] Schack, J. and Gawronski, J. (2008), "History does not repeat itself, it rhymes: The coming revolution in European market structure," *The Journal of Trading*, Fall, 71-81.